Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making

Kyle Dunovan[1,2†], **Catalina Vich**[3†], **Matthew Clapp**[4], **Timothy Verstynen**[1,2*¶], **Jonathan Rubin**[2,5*¶]

***For correspondence:**
timothyv@andrew.cmu.edu (TV);
jonrubin@pitt.edu (JR)

[†]These authors contributed equally to this work
[¶]These authors also contributed equally to this work

[1]Dept. of Psychology, Carnegie Mellon University; [2]Center for the Neural Basis of Cognition; [3]Dept. de Matemàtiques i Informàtica, Universitat de les Illes Balears; [4]Dept. of Biomedical Engineering, University of South Carolina; [5]Dept. of Mathematics, University of Pittsburgh

Abstract Cortico-basal-ganglia-thalamic (CBGT) networks are critical for adaptive decision-making, yet how changes to circuit-level properties impact cognitive algorithms remains unclear. Here we explore how dopaminergic plasticity at corticostriatal synapses alters competition between striatal pathways, impacting the evidence accumulation process during decision-making. Spike-timing dependent plasticity simulations showed that dopaminergic feedback based on rewards modified the ratio of direct and indirect corticostriatal weights within opposing action channels. Using the learned weight ratios in a full spiking CBGT network model, we simulated neural dynamics and decision outcomes in a reward-driven decision task and fit them with a drift-diffusion model. Fits revealed that the rate of evidence accumulation varied with inter-channel differences in direct pathway activity while boundary height varied with overall indirect pathway activity. This multi-level modeling approach demonstrates how complementary learning and decision computations emerge from corticostriatal plasticity.

Introduction

The flexibility of mammalian behavior showcases the dynamic range over which neural circuits can be modified by experience and the robustness of the emergent cognitive algorithms that guide goal-directed actions. Decades of research in cognitive science has independently detailed the algorithms of decision-making (e.g., accumulation-to-bound models, ***Ratcliff*** (***1978***)) and reinforcement learning (RL; ***Sutton et al.*** (***1998***); ***Rescorla et al.*** (***1972***)), providing foundational insights into the computational principles of adaptive decision-making. In parallel, research in neuroscience has shown how the selection of actions, and the use of feedback to modify selection processes, both rely on a common neural substrate: cortico-basal ganglia-thalamic (CBGT) circuits (***Doya, 2008***; ***Bogacz and Gurney, 2007***; ***Balleine et al., 2007***; ***Dunovan and Verstynen, 2016***).

Understanding how the cognitive algorithms for adaptive decision-making emerge from the circuit-level dynamics of CBGT pathways requires a careful mapping across levels of analysis (***Marr and Poggio, 1976***), from circuits to algorithm (see also ***Krakauer et al.*** (***2017***); ***Simen et al.*** (***2006***)). Previous simulation studies have demonstrated how the specific circuit-level computations of CBGT pathways map onto sub-components of the multiple sequential probability ratio test (MSPRT; ***Bo-***

gacz and Gurney (*2007*); *Bogacz* (*2007*)), a simple algorithm of information integration that selects single actions from a competing set of alternatives based on differences in input evidence (*Draglia et al., 1999*; *Baum and Veeravalli, 1994*). Allowing a simplified form of RL to modify corticostriatal synpatic weights results in an adaptive variant of the MSPRT that approximates the optimal solution to the action selection process based on both sensory signals and feedback learning (*Bogacz and Larsen, 2011*; *Caballero et al., 2018*). Previous attempts at multi-level modeling have largely adopted a "downwards mapping" approach, whereby the stepwise operations prescribed by computational or algorithmic models are intuitively mapped onto plausible neural substrates. Recently, *Frank* (*2015*) proposed an alternative "upwards mapping" approach for bridging levels of analysis, where biologically detailed models are used to simulate behavior that can be fit to a particular cognitive algorithm. Rather than ascribing different neural components with explicit computational roles, this variant of multi-level modeling examines how cognitive mechanisms are influenced by changes in the functional dynamics or connectivity of those components. A key assumption of the upwards mapping approach is that variability in the configuration of CBGT pathways should drive systematic changes in specific sub-components of the decision process, expressed by the parameters of the drift-diffusion model (DDM; *Ratcliff* (*1978*)). Indeed, by fitting the DDM to synthetic choice and response time data generated by a rate-based CBGT network, *Ratcliff and Frank* (*2012*) showed how variation in the height of the decision threshold tracked with changes in the strength of subthalamic nulceus (STN) activity. Thus, this example shows how simulations that map up the levels of analysis can be used to investigate the emergent changes in information processing that result from targeted modulation of the underlying neural circuitry.

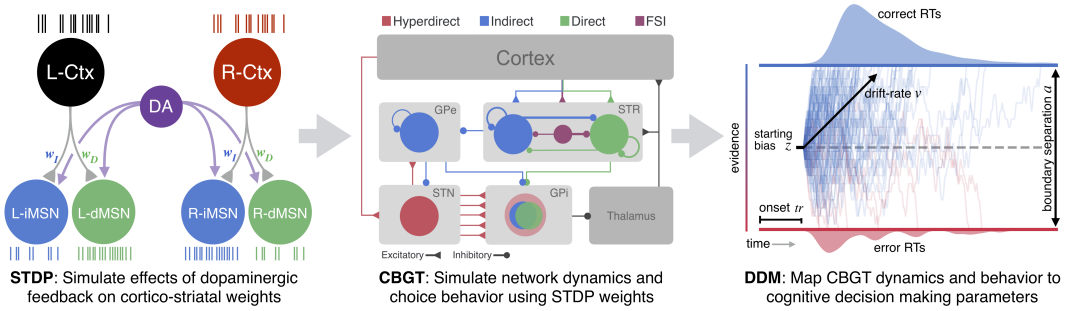


Figure 1. Multi-level modeling design. Left: An STDP model of DA effects on Ctx-dMSN and Ctx-iMSN synapses is used to determine how phasic DA signals affect the balance of these synapses. Middle: A spiking model of the CBGT pathways simulates behavioral responses, under different conditions of Ctx-MSN efficacy based on the STDP simulations. Right: The simulated behavioral responses from the full CBGT network model are then fit to a DDM of two-alternative choice behavior. Notation: $j-Ctx$ - cortical population, $j-dMSN$ - direct pathway striatal neurons, $j-iMSN$ - indirect pathway striatal neurons ($j \in \{L, R\}$); DA - dopamine signal; STR - striatum; GPe - globus pallidus external segment; STN - subthalamic nucleus; GPi - globus pallidus internal segment; FSI - fast spiking interneuron; RT - reaction time; v - DDM drift rate; a - separation between boundaries in DDM; z - bias in starting height of DDM; tr - time after which evidence accumulation begins in DDM.

Motivated by the predictions of a recently proposed Believer-Skeptic hypothesis of CBGT pathway function (*Dunovan and Verstynen, 2016*), we utilize the upwards mapping approach to modeling adaptive choice behavior across neural and cognitive levels of analysis (*Figure 1*). The Believer-Skeptic hypothesis posits that competition between the direct (Believer) and indirect (Skeptic) pathways within an action channel encodes the degree of uncertainty for that action. This competition is reflected in the drift rate of an accumulation-to-bound process (see *Dunovan et al.* (*2015*)). Over time, dopaminergic (DA) feedback signals can sculpt the Believer-Skeptic competition to bias decisions towards the behaviorally optimal target (*Bogacz and Larsen, 2011*). To explicitly test this prediction, we first modeled how phasic DA feedback signals (*Schultz et al., 1992*) can modulate the relative balance of corticostriatal synapses via spike-timing dependent plasticity (STDP; *Gurney et al.* (*2015*); *Baladron et al.* (*2017*)), thereby promoting or deterring action selection. The effects

of learning on the synaptic weights were subsequently implemented in a spiking model of the full CBGT network meant to accurately capture the known physiological properties and connectivity patterns of the constituent neurons in these circuits (***Wei et al., 2015***). The performance (i.e., accuracy and response times) of the CBGT simulations were then fit using a hierarchical DDM (***Wiecki and Frank, 2013***). This progression from synapses to networks to behavior, allows us to explicitly test the mechanistic predictions of the Believer-Skeptic hypothesis by mapping how specific features of striatal activity that result from reward-driven changes in corticostriatal synaptic weights could underlie parameters of the fundamental cognitive algorithms of decision-making.

Results

STDP network 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 artificial 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; *Action and rewards* subsection of *Methods*). These model MSNs were grouped into action channels receiving inputs from distinct cortical sources (***Figure 1***, left). Every time an action was selected, dopamine was released, after a short delay, at an intensity proportional to a reward prediction error (***Equation 9*** and ***Equation 10***). 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 (***Equation 8***; see Figure 10).

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, to assist in parameter tuning and verify effective model operation. In this scenario, where the rewards for each action did not change over time (i.e., one action always elicited a larger reward than the other), a gradual change in corticostriatal synaptic weights occurred (Supplementary ***Figure 1***A) in parallel with the learning of the actions' values (Supplementary ***Figure 1***B). These changes in synaptic weights induced altered MSN firing rates (Supplementary ***Figure 1***C,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 (***Figure 2***A). That is, firing rates in the direct pathway MSNs (dMSNs; D_L and D_R) associated with the more highly rewarded action increased, lead to a more frequent selection of that action. On the other hand, firing rates of the indirect pathway MSNs (iMSNs; I_L and I_R) remained quite similar (Supplementary ***Figure 1***C,D). This similarity is consistent with recent experimental results (***Donahue et al., 2018***), while the finding that dMSNs and iMSNs associated with a selected action are both active has also been reported in several experimental works (***Cui et al., 2013***; ***Tecuapetla et al., 2014***, ***2016***).

In this model, indirect pathway activity counters action selection by cancelling direct pathway spiking (*Action and rewards* subsection of *Methods*). This serves as a proxy in this simplified framework for indirect pathway competition with the direct pathway in the full network simulations (see *CBGT Dynamics and Choice Behavior* subsection of *Results*). Based on the cancellation framework, 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 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 (***Figure 2***).

To show that the network remained 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. Once values switched, the network was always able to learn the new values. Specifically, Q_L and Q_R

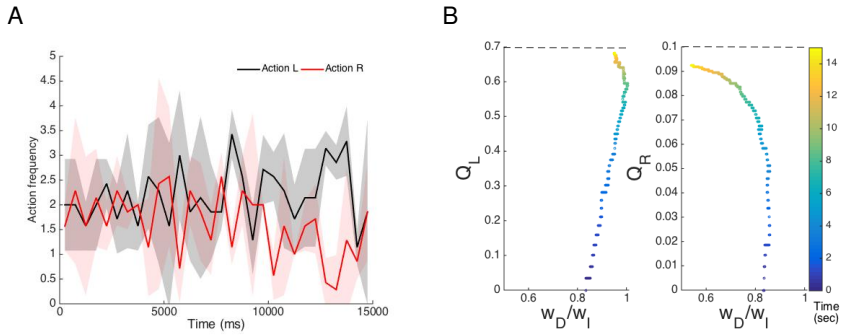


Figure 2. Constant reward task. A: Frequency of performance of L (black) and R (red) actions over time (discretized each $50\,ms$) when the rewards are held constant ($r_L = 0.7, r_R = 0.1$). Both traces are averaged across 7 different realizations. The transparent regions depict standard deviations. B: Estimates of the value of L (Q_L, left panel) and R (Q_R, right panel) versus the ratio of the corticostriatal weights to those dMSN neurons that facilitate the action and those iMSN that interfere with the action. Each trajectory is colored to show the progression of time. 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).

began evolving toward the new reward levels, switching their relative magnitudes along the way; the weights of corticostriatal synapses to L-dMSN (R-dMSN) weakened (strengthened) (e.g., ***Appendix 1 Figure 1***), and the relative performance frequencies of the two actions also reversed. Thus the network was able to adaptively learn immediate reward contingencies, without being restricted by previously learned contingencies.

While these 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 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 (***Frank et al., 2015***). 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 (***Figure 3***). 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 (***Figure 4***, 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 smaller values as the conflict 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

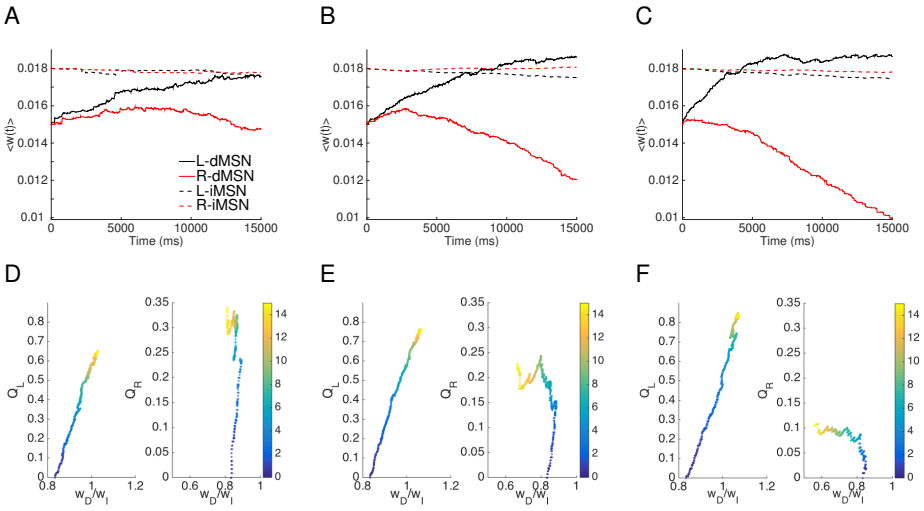


Figure 3. Corticostriatal synaptic weights when the reward traces are probabilistic. First column: $p_L = 0.65$; second column: $p_L = 0.75$; third column: $p_L = 0.85$ case. A, B, and C: Averaged weights over each of four specific populations of neurons, which are dMSN neurons selecting action L (solid 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, left panel) and R (Q_R, right panel) 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. Both the weights and the ratios have been averaged over 8 different realizations.

yield a reward (***Figure 4***F).

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 (***Figure 5***).

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 (***Frank et al., 2015***). 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 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 (***Figure 6***A) or averaged over all subjects experiencing the same reward probabilities (***Figure 6***B), as well as the probability of selection of action L averaged over subjects (***Figure 6***C). 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 were close to the actual values for many individual subjects as well as in mean (***Figure 6***A,B). These results agree quite well with the behavioral data in ***Frank et al.*** (***2015***) obtained from 15 human subjects, as well as with observations from similar experiments with rats (***Tort et al., 2009***). Also as in the experiments, the probability of selection of the more rewarded action grew across trials for all three reward probabilities, with less separation in action selection probability than in action values across different reward probability regimes (***Figure 6***C). Although our actual values for the probabilities of selection of higher value actions did not reach the levels seen experimentally, this likely reflected the non-biological action selection rule in our STDP model (see *Action and rewards* subsection of *Methods*), whereas the agreement of

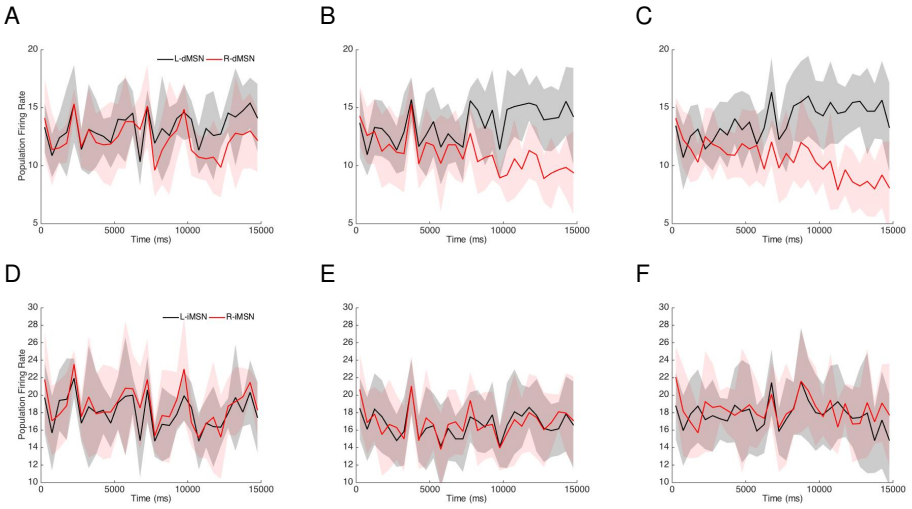


Figure 4. Firing rates when the reward traces are probabilistic. First column: $p_L = 0.65$; second column: $p_L = 0.75$; third column: $p_L = 0.85$ case. A, B and C: Time courses of firing rates of the dMSNs selecting the L (black) and R (red) actions ($50\,ms$ time discretization). D, E, and F: Time 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.

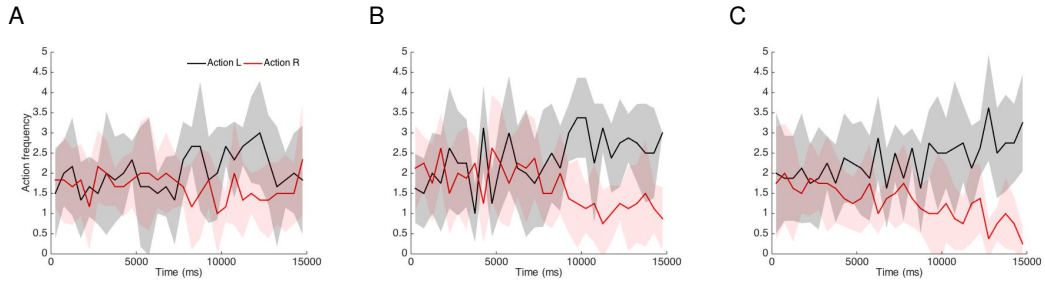


Figure 5. Action frequencies when reward delivery is probabilistic. All panels represent the number of L (black) and R (red) actions performed across time (discretized each $50\,ms$) when action selection is rewarded with probability $p_L = 0.65$ (A), $p_L = 0.75$ (B), or $p_L = 0.85$ (C) with $p_L + p_R = 1$. Traces represent the means over 8 different realizations, while the transparent regions depict standard deviations.

our model performance with experimental time courses of value estimation (***Figure 6***A,B) and our model's general success in learning to select more valuable actions (***Figure 1***C–Suppl. figure and ***Figure 5***) justify the incorporation of our results on corticostriatal synaptic weights into a spiking network with a more biologically-based decision-making mechanism, which we next discuss.

CBGT Dynamics and Choice Behavior

A key observation from our STDP model is that differences in rewards associated with different actions lead to differences in the ratios of corticostriatal synaptic weights to dMSN and iMSNs across action channels. Using weight ratios adapted from the STDP model, obtained by varying weights to dMSNs with fixed weights to iMSNs (Figure 3), we next performed simulations with a full spiking CBGT network to study the effects of this corticostriatal imbalance on the emergent neural dynamics and choice behavior following feedback-dependent learning in the context of low, medium, and high probability reward schedules (2500 trials/condition; see *Neural dynamics* subsection of *Methods* for details). In each simulation, cortical inputs featuring gradually increasing firing rates were supplied to both action channels, with identical statistical properties of inputs to both channels. These inputs led to evolving firing rates in nuclei throughout the basal ganglia,

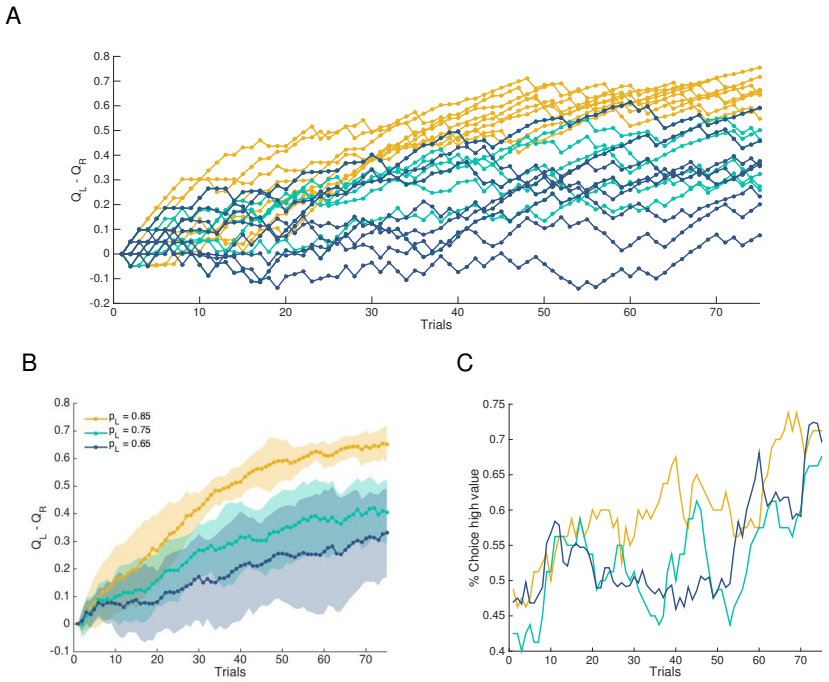


Figure 6. Relative action value estimates and action selection probabilities over simulated action selection trials with probabilistic reward schedules, with $p_L = 0.65$ (dark blue), $p_L = 0.75$ (cyan), $p_L = 0.85$ (yellow) and $p_L + p_R = 1$. A: Difference in action value estimates over trials in 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.

also partitioned into action channels, with an eventual action selection triggered by the thalamic firing rate in one channel reaching $30\,Hz$ (***Figure 1***, center and ***Figure 7***). We found that both dMSN and iMSN firing rates gradually increased in response to cortical inputs. Consistent with our STDP simulations (***Figure 4***), dMSN firing rates became higher in the channel for the selected action. Interestingly, iMSN firing rates also became higher in the selected channel, consistent with recent experiments (see ***Parker et al.*** (***2018***), among others). Similar to the activity patterns observed in the striatum, higher firing rates were also observed in the selected channel's STN and thalamic populations, whereas GPe and GPi firing rates were higher in the unselected channel (***Figure 7***).

More generally across all weight ratio conditions, dMSNs and iMSNs exhibited a gradual ramping in population firing rates (***Yartsev et al., 2018***) that eventually saturated around the average RT in each condition (***Figure 8***A). To characterize the relevant dimensions of striatal activity that contributed to the network's behavior, we extracted several summary measures of dMSN and iMSN activity, shown in ***Figure 8***B-C. Summary measures of dMSN and iMSN activity in the L and R channels were calculated by estimating the area under the curve (AUC) of the population firing rate between the time of stimulus onset ($200\,ms$) and the RT on each trial. Trialwise AUC estimates were then normalized between values of 0 and 1, including estimates from all trials in all conditions in the normalization. As expected, increasing the disparity of left and right Ctx-dMSN weights led to greater differences in direct pathway activation between the two channels (i.e., $D_L > D_R$; ***Figure 8***B). The increase in $D_L - D_R$ reflects a form of competition *between* action channels, where larger values indicate stronger dMSN activation in the optimal channel and/or a weakening of dMSN activity in the suboptimal channel. Similarly, increasing the weight of Ctx-dMSN connections caused a shift in the competition between dMSN and iMSN populations *within* the left action channel (i.e., $D_L > I_L$). Thus, manipulating the weight of Ctx-dMSN connections to match those predicted by the STDP model led to both between- and within-channel biases favoring firing of the direct pathway of the

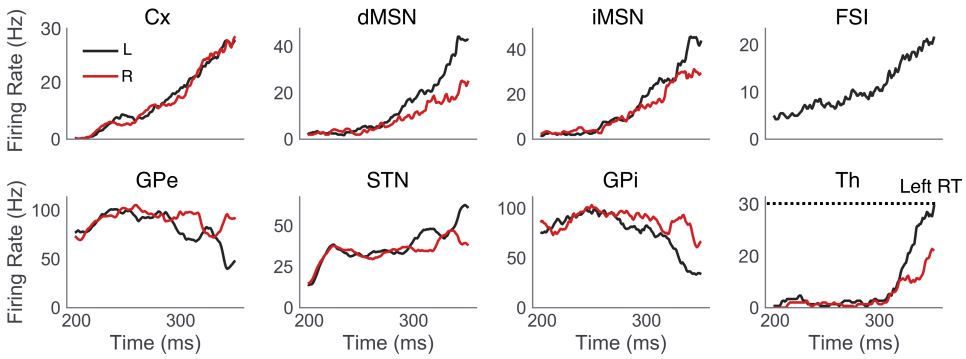


Figure 7. Single trial example of CBGT dynamics. Population firing rates of CBGT nuclei, computed as the average of individual unit firing rates within each nucleus in L (black) and R (red) action channels are shown for a single representative trial in the high reward probability condition. The selected action (L) and corresponding RT ($324\,ms$) are determined by the first action channel to raise its thalamic firing rate to $30\,Hz$.

optimal action channel in proportion to its expected reward value.

Interestingly, although the weights of Ctx-iMSN connections were kept constant across conditions, iMSN populations showed reliable differences in activation between channels (***Figure 8***C). Similar to the observed effects on direct pathway activation, higher reward conditions were associated with progressively greater differences in the AUC of L and R indirect pathway firing rates ($I_L - I_R$). At first glance, greater indirect pathway activation in higher compared to lower valued action channels differs from the similarity of activation levels of both indirect pathway channels that we obtained in the STDP model and also appears to be at odds with canonical theories of the roles of the direct and indirect pathways in RL and decision-making. This finding can be explained, however, based on a certain feature represented in the connections within the CBGT network but not within the STDP network, namely thalamo-striatal feedback between channels. That is, the strengthening and weakening of Ctx-dMSN weights in the L and R channels, respectively, translated into relatively greater downstream disinhibition of the thalamus in the L channel, which increased excitatory feedback to L-dMSNs and L-iMSNs while reducing thalamo-striatal feedback to R-MSNs in both pathways.

Finally, we examined the effects of reward probability on the AUC of all iMSN firing rates (I_{all}; combining across action channels). Observed differences in I_{all} across reward conditions were notably more subtle than those observed for other summary measures of striatal activity, with greatest activity in the medium reward condition, followed by the high and low reward conditions, respectively.

In addition to analyzing the effects of altered Ctx-dMSN connectivity strength on the functional dynamics of the CBGT network, we also studied how the decision-making behavior of the CBGT network was influenced by this manipulation. Consistent with previous studies of value-based decision-making in humans (***Manohar et al., 2015***; ***Polanía et al., 2014***; ***Afacan-Seref et al., 2018***; ***Gardner et al., 2017***; ***Jahfari et al., 2017***), we observed a positive effect of reward probability on both the frequency and speed of correct (e.g., leftward, associated with higher reward probability) choices (***Figure 8***D). Bootstrap sampling (10,000 samples) was performed to estimate 95% confidence intervals (CI_{95}) around RT and accuracy means (μ) in each condition, and to assess the statistical significance of pairwise comparisons between conditions. Choice accuracy increased across low ($\mu = 64\%, CI_{95} = [62, 65]$), medium ($\mu = 85\%, CI_{95} = [84, 86]$), and high ($\mu = 100\%, CI_{95} = [100, 100]$) reward probabilities. Pairwise comparisons revealed that the increase in accuracy observed between low and medium conditions, as well as that observed between medium and high conditions, reached statistical significance (both $p < 0.0001$). Along with the increase in accuracy across

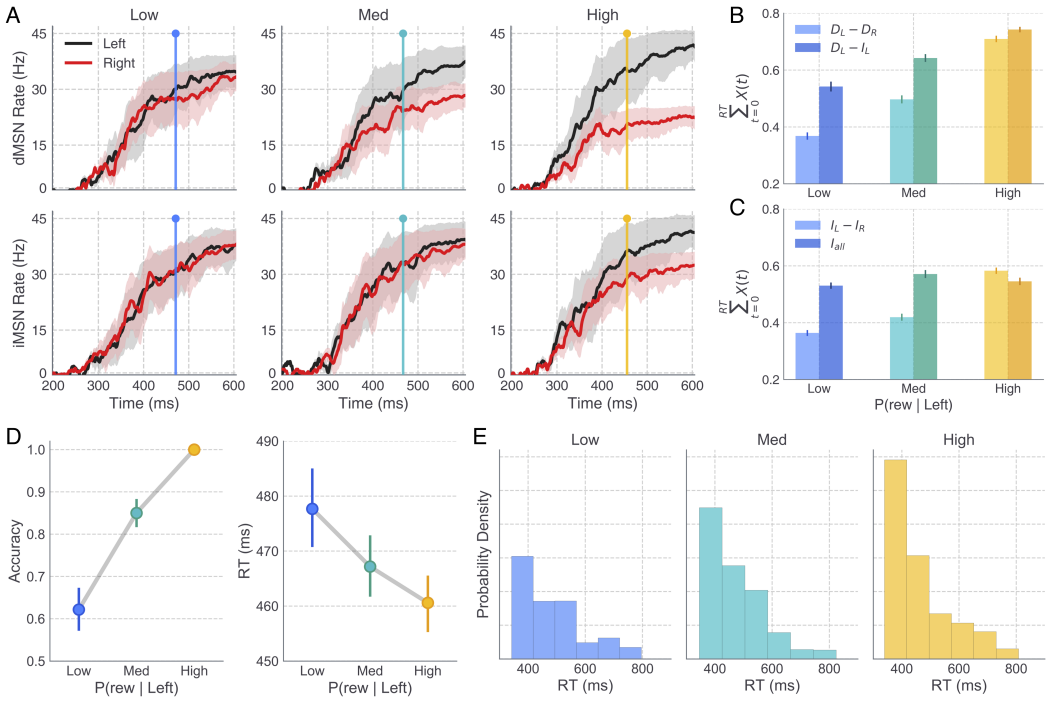


Figure 8. Striatal pathway dynamics and behavioral effects of reward probability in full CBGT network. A: Time courses show the average population firing rates for L (black) and R (red) dMSNs (top) and iMSNs (bottom) over the the trial window. Shaded areas reflect 95% CI. Colored vertical lines depict the average RT in the low (blue), medium (cyan), and high (yellow) reward conditions. B and C: Summary statistics of dMSN and iMSN population firing rates were extracted on each trial and later included as trial-wise regressors on parameters of the DDM, allowing specific hypotheses to be tested about the mapping between neural and cognitive mechanisms. In B, lighter colored bars show the difference between dMSN firing rates in the L and R action channels whereas darker colored bars show the difference between dMSN and iMSN firing rates in the L action channel, both computed by summing the average firing rate of each population between trial onset and the RT on each trial. In C, lighter colored bars show the difference between iMSN firing rates in the L and R action channels and darker colored bars show the average iMSN firing rate (combined across left and right channels). Error bars show the bootstrapped 95% CI. D: Average accuracy (probability of choosing L) and RT (L choices only) of CBGT choices across levels of reward probability. E: RT distributions for correct choices across levels of reward probability; note that higher reward yields more correct trials. Error bars in B-D show the bootstrapped 95% CI.

conditions, we observed a concurrent decrease in the RT of correct (L) choices in the low ($\mu = 477\text{ms}, \text{CI}_{95} = [472, 483]$), medium ($\mu = 467\text{ms}, \text{CI}_{95} = [462, 471]$), and high ($\mu = 460\text{ms}, \text{CI}_{95} = [456, 464]$) reward probability conditions. Notably, our manipulation of Ctx-dMSN weights across conditions manifested in stronger effects on accuracy (i.e., probability of choosing the more valuable action), with subtler effects on RT. Specifically, the decrease in RT observed between the low and medium conditions reached statistical significance ($p < .0001$); however, the RT decrease observed between the medium and high conditions did not ($p = .13$).

We also examined the distribution of RTs for L responses across reward conditions (***Figure 8***E). All conditions showed a rightward skew in the distribution of RTs, an empirical hallmark of simple choice behavior and a useful check of the suitability of accumulation-to-bound models like the DDM for modeling a particular behavioral data set. Moreover, the degree of skew in the RT distributions for L responses became more pronounced with increasing reward probability, suggesting that the observed decrease in the mean RT at higher levels of reward was driven by a change in the shape of the distribution, and not, for instance, a temporal shift in its location.

CBGT-DDM Mapping

We performed fits of a normative DDM to the CBGT network's decision-making performance (i.e., accuracy and RT data) to understand the effects of corticostriatal plasticity on emergent changes in decision behavior. This process was implemented in three stages. First, we compared models in which only one free DDM parameter was allowed to vary across levels of reward probability (single parameter DDMs). Next, a second round of fits was performed in which a second free DDM parameter was included in the best-fitting single parameter model identified in the previous stage (dual parameter DDMs). Finally, the two best-fitting dual parameter models were submitted to a third and final round of fits with the inclusion of trialwise measures of striatal activity (see ***Figure 8***B-C) as regressors on designated parameters of the DDM.

All models were evaluated according to their relative improvement in performance compared to a null model in which all parameters were fixed across conditions. To identify which single parameter of the DDM best captured the behavioral effects of alterations in reward probability as represented by Ctx-dMSN connectivity strength, we compared the deviance information criterion (DIC) of models in which either the boundary height (a), the onset delay (tr), the drift rate (v), or the starting-point bias (z) was allowed to vary across conditions. ***Figure 9***A shows the difference between the DIC score of each model (DIC_M) and that of the null model ($\Delta\text{DIC} = \text{DIC}_M - \text{DIC}_{null}$), with lower values indicating a better fit to the data (see ***Table 1*** for additional fit statistics). Conventionally, a DIC difference (ΔDIC) of magnitude 10 or more is regarded as strong evidence in favor of the model with the lower DIC value (***Burnham and Anderson, 1998***). Compared to the null model as well as alternative single parameter models, allowing the drift rate v to vary across conditions afforded a significantly better fit to the data ($\Delta\text{DIC} = -960.79$). Examination of posterior distributions of v in the best-fitting single parameter model revealed a significant increase in v with successively higher levels of reward probability ($v_{Low} = .35$; $v_{Med} = 1.61$; $v_{High} = 2.71$), capturing the observed increase in speed and accuracy across conditions by increasing the rate of evidence accumulation toward the upper (L) decision threshold.

To investigate potential interactions between the drift rate and other parameters of the DDM, we performed another round of fits in which a second free parameter (either a, tr, or z), in addition to v, was allowed to vary across conditions (***Figure 9***A). Compared to alternative dual-parameter models, the combined effect of allowing v and a to vary across conditions (Fig. ***Figure 8***B,C) provided the greatest improvement in model fit over the null model ($\Delta\text{DIC} = -1174.07$), as well as over the best-fitting single parameter model ($\text{DIC}_{v,a} - \text{DIC}_v = -213.27$). While the dual v and a model significantly outperformed both alternatives ($\text{DIC}_{v,a} - \text{DIC}_{v,t} = -205.89$; $\text{DIC}_{v,a} - \text{DIC}_{v,z} = -184.05$), the second best-fitting dual parameter model, in which v and z were left free across conditions, also afforded a significant improvement over the drift-only model ($\text{DIC}_{v,z} - \text{DIC}_v = -29.23$). Thus, both v, a and v, z dual parameter models were considered in a third and final round of fits. The third round was motivated by the fact that, while behavioral fits can yield reliable and informative insights about the cognitive mechanisms engaged by a given experimental manipulation, recent studies have effectively combined behavioral observations with coincident measures of neural activity to test more precise hypotheses about the neural dynamics involved in regulating different cognitive mechanisms (***Herz et al., 2016***, ***2018***; ***Frank et al., 2015***). To this end, we refit the v, a and v, z models to the same simulated behavioral dataset (i.e., accuracy and RTs produced by the CBGT network) as in the previous rounds, with the addition of different trialwise measures of striatal activity included as regressors on one of the two free parameters in the DDM.

For each regression DDM (N=24 models, corresponding to 24 ways to map 2 of 6 striatal activity measures to the v, a and v, z models), one of the summary measures shown in ***Figure 8***B-C was regressed on v, and another regressed on either a or z, with separate regression weights estimated for each level of reward probability. Model fit statistics are shown for each of the 24 regression models in ***Table 2***, along with information about the neural regressors included in each model and their respective parameter dependencies. The relative goodness-of-fit afforded by all 24 regression

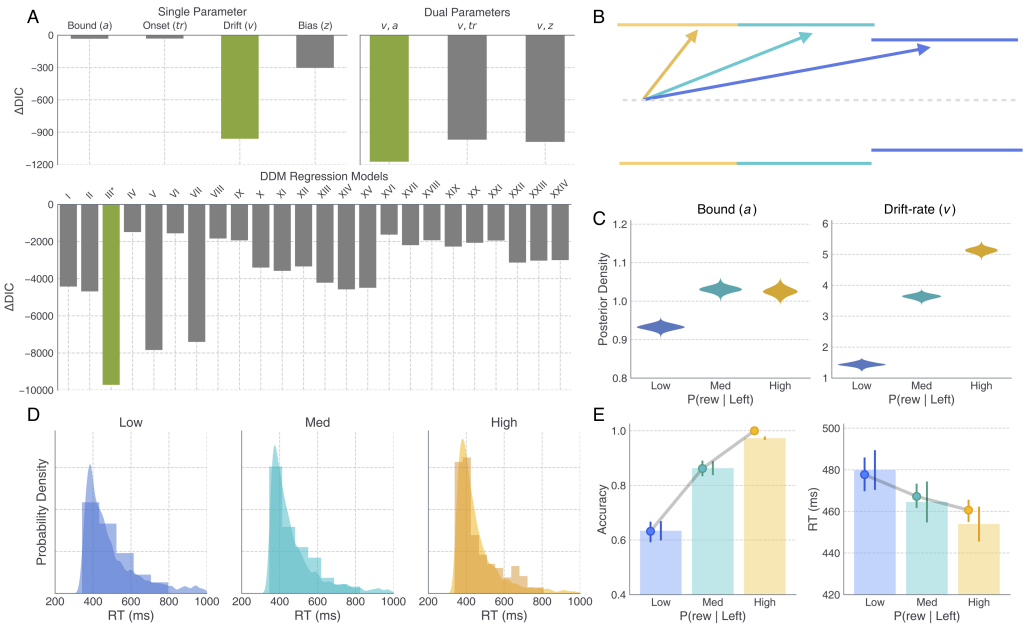


Figure 9. DDM fits to CBGT-simulated behavior reveals pathway-specific effects on drift rate and threshold mechanisms. A: ΔDIC scores, showing the relative goodness-of-fit of all single- and dual-parameter DDMs considered (top) and all DDM regression models considered (bottom) compared to that of the null model (all parameters held constant across conditions; see ***Table 2***). The ΔDIC score of the best-fitting model at each stage is plotted in green. The best overall fit was provided by DDM regression model III. B: DDM schematic showing the change in v and a across low (blue), medium (cyan), and high (yellow) reward conditions, with the threshold for L and R represented as the upper and lower boundaries, respectively. C: Posterior distributions showing the estimated weights for neural regressors on a, which was estimated on each trial as a function of the average iMSN firing rate across left and right action channels (see I_{all} in ***Figure 8***C), and v, which was estimated on each trial as a function of the the difference between dMSN firing rates in the left and right channels (see $D_L - D_R$ in ***Figure 8***B). D: Histograms and kernel density estimates showing the CBGT-simulated and DDM-predicted RT distributions, respectively. E: Point plots showing the CBGT network's average accuracy and RT across reward conditions overlaid on bars showing the DDM-predicted averages.

models is visualized in ***Figure 9***A (lower panel), identifying what we have labelled as model III as the clear winner with an overall DIC $= -18860.37$ and with $\Delta\text{DIC} = -9716.17$ compared to the null model. In model III, the drift rate v on each action selection trial depended on the relative strength of direct pathway activation in L and R action channels (e.g., $D_L - D_R$), whereas the boundary height a on that trial was computed as a function of the overall strength of indirect pathway activation across both channels (e.g., I_{all}). To determine how these parameter dependencies influenced levels of v and a across levels of reward probability, the following equations were used to transform intercept and regression coefficient posteriors into posterior estimates of v and a for each condition j:

$$v_j = \beta_0^v + \beta_j^v \Delta D_j, \tag{1}$$

$$a_j = \beta_0^a + \beta_j^a I_j, \tag{2}$$

where ΔD_j and I_j are the mean values of $D_L - D_R$ and I_{all} in condition j (see ***Figure 8***B-C), β_0^v and β_0^a are posterior distributions for v and a intercept terms, and β_j^v and β_j^a are the posterior distributions estimated for the linear weights relating $D_L - D_R$ and I_{all} to v and a, respectively. The observed effects of reward probability on v and a, as mediated by trialwise changes in $D_L - D_R$ and I_{all}, are schematized in ***Figure 9***B, with conditional posteriors for each parameter plotted in ***Figure 9***C. Consistent with best-fitting single and dual parameter models (e.g., without striatal regressors included), the weighted effect of $D_L - D_R$ on v in model III led to a significant increase in v across

low ($\mu_{v_{Low}} = 1.43, \sigma_{v_{Low}} = .063$), medium ($\mu_{v_{Med}} = 3.62, \sigma_{v_{Med}} = .078$), and high ($\mu_{v_{High}} = 5.10, \sigma_{v_{High}} = .086$) conditions. Thus, increasing the disparity of dMSN activation between L and R action channels led to faster and more frequent leftward actions by increasing the rate of evidence accumulation towards the correct decision boundary. Also consistent with parameter estimates from the best-fitting dual parameter model (i.e., v, a), inclusion of trialwise values of I_{all} led to an increase in the boundary height in the medium ($\mu_{a_{Med}} = 1.025, \sigma_{a_{Med}} = .009$) and high ($\mu_{a_{High}} = 1.020, \sigma_{a_{High}} = .011$) conditions compared to estimates in the low condition ($\mu_{a_{Low}} = 0.93, \sigma_{a_{Low}} = .008$). However, in contrast with boundary height estimates derived from behavioral data alone (not shown), a estimates in model III showed no significant difference between medium and high levels of reward probability.

Next, we evaluated the extent to which the best-fitting regression model (i.e., model III) was able to account for the qualitative behavioral patterns exhibited by the CBGT network in each condition. To this end, we simulated 20,000 trials in each reward condition (each trial producing a response and RT given a parameter set sampled from the model posteriors) and compared the resulting RT distributions, along with mean speed and accuracy measures, with those produced by the CBGT model (***Figure 9***D,E). Parameter estimates from the best-fitting model captured both the increasing rightward skew of RT distributions, as well as the concurrent increase in mean decision speed and accuracy with increasing reward probability.

In summary, by leveraging trialwise measures of simulated striatal MSN subpopulation dynamics to supplement RT and choice data generated by the CBGT network, we were able to 1) substantially improve the quality of DDM fits to the network's behavior across levels of reward probability compared to models without access to neural observations and 2) identify dissociable neural signals underlying observed changes in v and a across varying levels of reward probability associated with available choices.

Table 1. Single- and dual-parameter DDM goodness-of-fit statistics. DIC is a complexity-penalized measure of model fit, DIC = D(θ) + pD, where D(θ) is the deviance of model fit under the optimized parameter set θ and pD is the effective number of parameters. ΔDIC is the difference between each model's DIC and that of the null model for which all parameters are fixed across conditions. Asterisks denote models providing best fits within the single-parameter group (*) and across both groups (**).

	DIC	ΔDIC
Null	-9144.21	0.0
Bound (a)	-9177.03	-32.83
Onset (tr)	-9175.54	-31.34
Drift (v)*	-10105.0	-960.79
Bias (z)	-9447.5	-303.29
v, a**	-10318.27	-1174.07
v, tr	-10113.38	-969.17
v, z	-10134.22	-990.02

Discussion

Reinforcement learning in mammals alters the mapping from sensory evidence to action decisions. Here we set out to understand how this adaptive decision-making process emerges from underlying neural circuits using a modeling approach that bridges across levels of analysis, from plasticity at corticostriatal synapses to CBGT network function to quantifiable behavioral parameters (***Simen et al., 2006***; ***Ratcliff and Frank, 2012***; ***Bogacz, 2007***; ***Bogacz and Larsen, 2011***). We show how a simple, DA-mediated STDP rule can modulate the sensitivity of both dMSN and iMSN populations to cortical inputs. This learning allows for the network to discover which target in a two-alternative forced-choice task is more likely to deliver a reward by modifying the ratio of direct and indirect

Table 2. DDM regression models and goodness-of-fit statistics. Asterisk denotes best performing model.

	$D_L - D_R$	$D_L - I_L$	$I_L - I_R$	I_{all}	DIC	ΔDIC
I	v	a	–	–	-13567.84	-4423.64
II	v	–	a	–	-13828.38	-4684.17
***III**	v	–	–	a	-18860.37	-9716.16
IV	–	v	a	–	-10636.70	-1492.50
V	–	v	–	a	-16982.35	-7838.14
VI	a	v	–	–	-10702.48	-1558.27
VII	–	–	v	a	-16547.47	-7403.27
VIII	a	–	v	–	-10979.51	-1835.31
IX	–	a	v	–	-11082.55	-1938.34
X	a	–	–	v	-12546.90	-3402.70
XI	–	a	–	v	-12719.92	-3575.72
XII	–	–	a	v	-12486.66	-3342.46
XIII	v	z	–	–	-13361.52	-4217.32
XIV	v	–	z	–	-13719.36	-4575.16
XV	v	–	–	z	-13634.12	-4489.92
XVI	–	v	z	–	-10774.88	-1630.67
XVII	–	v	–	z	-11340.47	-2196.26
XVIII	z	v	–	–	-11074.84	-1930.64
XIX	–	–	v	z	-11418.76	-2274.56
XX	z	–	v	–	-11213.79	-2069.59
XXI	–	z	v	–	-11090.96	-1946.75
XXII	z	–	–	v	-12279.57	-3135.36
XXIII	–	z	–	v	-12171.17	-3026.96
XXIV	–	–	z	v	-12144.98	-3000.77

pathway corticostriatal weights within each action channel. With this result in hand, we simulated the network-level dynamics of CBGT circuits, as well as behavioral responses, under different levels of conflict in reward probabilities, by extrapolating from the learned corticostriatal weights from the STDP simulations. As reward probability for the optimal target increased, the asymmetry of dMSN firing rates between action channels grew, as did the overall activity of iMSNs across both action channels. By fitting the DDM to the simulated decision behavior of the CBGT network, we found that changes in the rate of evidence accumulation tracked with the difference in dMSN population firing rates across action channels, while the the level of evidence required to trigger a decision tracked with the overall iMSN population activity. These findings show how, at least within this specific framework, plasticity at corticostriatal synapses induced by phasic changes in DA can have a multifaceted effect on cognitive decision processes.

A critical assumption of our theoretical experiments is that the CBGT pathways accumulate sensory evidence for competing actions in order to identify the most contextually appropriate response. This assumption is supported by a growing body of empirical and theoretical evidence. For example, ***Yartsev et al.*** (***2018***) recently showed that, in rodents performing an auditory discrimination task, the anterior dorsolateral striatum satisfied three fundamental criteria for establishing causality in the evidence accumulation process: (1) inactivation of the striatum ablated the animal's discrimination performance on the task, (2) perturbation of striatal neurons during the temporal window of evidence accumulation had predictable and reliable effects on trial-wise behavioral reports, and (3) gradual ramping, proportional to the strength of evidence, was observed in both single unit and population firing rates of the striatum (however, see also ***Ding and Gold*** (***2010***)). Consistent with these empirical findings, ***Caballero et al.*** (***2018***) recently proposed a novel computational frame-

work, capturing perceptual evidence accumulation as an emergent effect of recurrent activation of competing action channels. This modeling work builds on previous studies showing how the architecture of CBGT loops is ideal for implementing a variant of the sequential probability ratio test (***Bogacz and Gurney, 2007***; ***Bogacz, 2007***). Taken together, these converging lines of evidence point to CBGT pathways as being causally involved in the accumulation of evidence for decision-making.

The idea that an accumulation of evidence algorithm can be implemented via network-level dynamics within looped circuit architectures stands in sharp contrast to cortical models of decision-making that presume a more direct isomorphism between accumulators and neural activity (for review see ***Gold and Shadlen*** (***2007***)). Early experimental work showed how population-level firing rates in area LIP displayed the same ramp-to-threshold dynamics as predicted by an evidence accumulation process (***Shadlen and Newsome, 2001***; ***Kiani and Shadlen, 2009***; ***Churchland et al., 2008***). This simple relation between algorithm and implementation has now come into question. Follow-up electrophysiological experiments showed how this population-level accumulation may, in fact, reflect the aggregation of step-functions across neurons that resemble an accumulator when summed together yet lack accumulation properties at the level of individual units (***Latimer et al., 2015***). In addition, recent results from intervention studies are inconsistent with the causal role of cortical areas in the accumulation of evidence. For instance, ***Katz et al.*** (***2016***) found that inactivation of area LIP in macaques had no effect on the ability of monkeys to discriminate the direction of motion stimuli in a standard random dot motion task. In contrast to the presumed centrality of LIP in sensory evidence accumulation, these findings and supporting reports from ***Licata et al.*** (***2017***) and ***Erlich et al.*** (***2015***) suggest that cortical areas like LIP provide a useful proxy for the deliberation process but are unlikely to have a causal role in the decision itself.

The recent experimental (***Yartsev et al., 2018***) and theoretical (***Caballero et al., 2018***) revelations of CBGT involvement in decision-making are particularly exciting, not only for the purposes of identifying a likely neural substrate of perceptual choice, but also for their implications for integrating accumulation-to-bound models (e.g., action selection mechanisms) with theories of RL (e.g., feedback-dependent learning of action values). We previously proposed a Believer-Skeptic framework (***Dunovan and Verstynen, 2016***) to capture the complementary roles played by the direct and indirect pathways in the feedback-dependent learning and the moment-to-moment evidence accumulation leading up to action selection. This competition between opposing control pathways can be characterized as a debate between a Believer (direct pathway) and a Skeptic (indirect pathway), reflecting the instantaneous probability ratio of evidence in favor of executing and suppressing a given action respectively. Because the default state of the basal ganglia pathways is motor-suppressing (e.g., ***Alexander and Crutcher*** (***1990***); ***Wichmann and DeLong*** (***1996***)), the burden of proof falls on the Believer to present sufficient evidence for selecting a particular action. In accumulation-to-bound models like the DDM, this sequential sampling of evidence is parameterized by the drift rate. Therefore, the Believer-Skeptic model specifically predicts that this competition should be reflected, at least in part, in the rate of evidence accumulation. As for the role of learning in the Believer-Skeptic competition, multiple lines of evidence suggest that dopaminergic feedback during learning systematically biases the direct-indirect competition in a manner consistent with increasing the drift rate for more rewarding actions (***Pedersen et al., 2017***; ***Manohar et al., 2015***; ***Collins and Frank, 2014***; ***Dunovan and Verstynen, 2016***; ***Frank et al., 2015***; ***Afacan-Seref et al., 2018***). Indeed, the STDP simulations in the current study showed opposing effects of dopaminergic feedback on corticostriatal synapses in the direct pathway for both the optimal and suboptimal action channels, with the post-learning difference between the direct pathway synaptic weights in the two channels proportional to the difference in expected action values. This provides testable predictions at multiple levels for how feedback learning should influence the decision process over time.

In support of the biological assumptions underlying the CBGT network, several important empirical properties naturally emerged from our simulations. First, both dMSN and iMSN striatal populations were concurrently activated on each trial (see ***Cui et al.*** (***2015***); ***Klaus et al.*** (***2017***);

Donahue et al. (***2018***)) and exhibited gradually ramping firing rates that often saturated before the response on each trial (***Yartsev et al., 2018***; ***Ding and Gold, 2010***). Second, in contrast with the relatively early onset of ramping activity in the striatum, recipient populations in the GPi sustained high tonic firing rates throughout most of the trial, with activity in the selected channel showing a precipitous decline near the recorded RT (***Schmidt et al., 2013***; ***Lo and Wang, 2006***; ***Wei et al., 2015***). This delayed change in GPi activation is likely due to the opposing influence of concurrently active dMSN and iMSN populations in each channel, such that the influence of the direct pathway on the GPi is temporarily balanced out by activation of the indirect pathway (see ***Wei et al.*** (***2015***)). To represent low, medium, and high levels of reward probability conflict, we manipulated the weights of cortical input to dMSNs in each channel (see ***Table 4***), increasing and decreasing the ratio of direct pathway weights to indirect pathway weights for L and R actions, respectively. As expected, increasing the difference in the associated reward for L and R actions led to stronger firing in L-dMSNs and weaker firing of R-dMSNs. Consistent with recently reported electrophysiological findings (***Donahue et al., 2018***; ***Klaus et al., 2017***), we also observed an increase in the firing of iMSNs in the L action channel, which in our simulations may arise from channel-specific feedback from the L component of the thalamus. Behaviorally, the choices of the CBGT network became both faster and more accurate (e.g., higher percentage of L responses) at higher levels of reward, suggesting that the observed increase in L-iMSN firing did not serve to delay or suppress L selections. These changes in neural dynamics also produced consequent changes in value-based decision behavior consistent with previous studies linking parameters of the DDM with experiential feedback .

One of the critical outcomes of the current set of experiments is the mechanistic prediction of how variation in specific neural parameters relates to changes in parameters of the DDM. Consistent with past work (see ***Dunovan and Verstynen*** (***2016***); ***Frank et al.*** (***2015***)), the DDM fits to the CBGT-simulated behavior showed an increase in drift rate toward the higher valued decision boundary with increasing expected reward. Additionally, we found that greater disparity in the expected values of alternative actions led to an increase in the boundary height. Indeed, the co-modulation of drift rate and boundary parameters observed here has also been found in human and animal experimental studies of value-based choice (***Afacan-Seref et al., 2018***; ***Manohar et al., 2015***; ***Frank et al., 2015***). For example, experiments with human subjects in a value-based learning task showed that selection and response speed patterns were best described by an increase in the rate of evidence for more valued targets, coupled with an upwards shift in the boundary height for all targets (***Manohar et al., 2015***). Moreover, in healthy human subjects, but not Parkinson's disease patients, reward feedback was found to drive increases in both rate and boundary height parameters, effectively breaking the speed-accuracy tradeoff ***Manohar et al.*** (***2015***). To identify more precise links between the relevant neural dynamics underlying the observed drift rate and boundary height effects we performed another round of model fits with striatal summary measures included as regressors to describe trial-by-trial variability. Behavioral fits were substantially improved by estimating trialwise values of drift rate as a function of the difference between L- and R-dMSN activation and trialwise values of boundary height as a function of the iMSN activation across both channels. These relationships stand both as novel predictions arising from the current study and as refinements to the Believer-Skeptic framework, implying that the Believer component relies on a competition between action channels while the Skeptic involves a cooperative aspect.

While our present findings provide key insights into the links between implementation mechanisms and cognitive algorithms during adaptive decision-making, they are constrained by the nature of the multi-level modeling approach itself. When considering the potential benefits of multi-level modeling to investigations mapping across levels of analysis, it is important to understand the kinds of questions that this approach is poised to address as well as some of its limitations. For instance, because the CBGT network is substantially more complex (i.e., has more parameters) than the DDM, it is necessarily more flexible in terms of the empirical phenomena that it is capable of fitting.

Thus, given this disparity in the degrees of freedom at the neural and cognitive levels, it becomes exceedingly likely that multiple parameter mappings exist between them. That is, many different properties of the CBGT network, aside from corticostriatial weights and measures of striatal activity, could potentially be manipulated to cause analogous behavioral patterns and inferred effects on the drift rate and boundary height parameters in the DDM. Importantly, our goal in using multi-level modeling was not to arbitrate between alternative, mutually exclusive biological implementations of a given parameter that appears at the cognitive level, but rather to understand how specific neural features related to the Believer-Skeptic framework contribute to cognitive computations by examining which parameters they influence when perturbed. We do not presume that the impacts of dopaminergic plasticity at corticostriatal synapses on striatal activity are singularly responsible for setting the drift rate during value-based decision-making, for example. Rather, our simulations demonstrate that strengthening corticostriatal synapses is one way that the brain can adjust striatal firing to shape the drift rate and accumulation threshold, promoting faster and more frequent selection of actions with a higher expected value.

Our simulations make several novel predictions for future experiments. The STDP simulations described in the *STDP network results* section suggest that feedback-dependent reward learning should drive more salient changes in cortical synaptic weights to dMSN populations than to iMSN populations. At the same time, while the learning-related changes in L and R direct pathway corticostriatal weights were mirrored by the relative firing rates of L- and R-dMSNs in the CBGT network, iMSN firing rates are also predicted to show channel-specific differences, despite constancy in their corticostriatal weights across conditions. The observed increase in iMSN firing disparity between the L and R channels in our simulations emerged due to the thalamostriatal feedback assumed in the CBGT network, where dMSN activation leads to disinhibition of the thalamus, thereby increasing excitatory feedback to both MSN subtypes within a given channel. This represents another novel model prediction that can be tested empirically. Since it is currently unclear whether these feedback connections actually adhere to a channel-specific (e.g., focal) topology, we hope that our work will motivate future experiments to explore the topology of thalamostriatal inputs. Finally, our study predicts that the difference in dMSN activity across action channels modulates the rate of value-based evidence accumulation. This could be directly tested by applying different magnitudes of optogenetic stimulation to dMSNs in L- and R-lateralized dorsolateral striatum to effectively manipulate the strength of evidence for L and R lever presses. According to our simulations, increasing the relative magnitude of dMSN stimulation in the R, compared to L, dorsolateral striatum should speed and facilitate the selection of contralateral lever presses. Choice and RT data could then be fit with the DDM to determine if the behavioral effects of laterally-biased dMSN stimulation were best described by a change in the drift rate. Analogous experiments targeting iMSNs but without channel specificity could be used similarly to evaluate our prediction that overall iMSN activity level modulates DDM boundary height.

Conclusion

Here we characterize the effects of dopaminergic feedback on the competition between direct and indirect CBGT pathways and how this plasticity impacts the evaluation of evidence for alternative actions during value-based choice. Using simulated neural dynamics to generate behavioral data for fitting by the DDM and determining how measures of striatal activity influence this fit, we show how the rate of evidence accumulation and the decision boundary height are modulated by the direct and indirect pathways, respectively. This multi-level modeling approach affords a unique combination of biological plausibility and mechanistic interpretability, providing a rich set of testable predictions for guiding future experimental work at multiple levels of analysis.

Methods

Our work involves three distinct model systems, a *spike-timing dependent plasticity (STDP)* network consisting of striatal neurons and their cortical inputs, with corticostriatal synaptic plasticity driven

by phasic reward signals resulting from simulated actions and their consequent dopamine release; a spiking *cortico-basal ganglia-thalamic (CBGT)* network, comprising neurons and synaptic connections from the key cortical and subcortical areas within the CBGT computational loops that take sensory evidence from cortex and make a decision to select one of two available responses; and the *drift diffusion model (DDM)*, a cognitive model of decision-making that describes the accumulation-to-bound dynamics underlying the speed and accuracy of simple choice behavior (***Ratcliff, 1978***).

In this section, we present the details of each of these models along with some computational approaches that we use in simulating and analyzing them. The three models are simulated separately, but outputs of specific models are critical for the tuning of other models, as we shall describe.

STDP network

Neural model

We consider a computational model of the striatum consisting of two different populations that receive different inputs from the cortex (see ***Figure 1***, left). Although they do not interact directly, they 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 (***Fourcaud-Trocmé et al., 2003***), such that each neural membrane potential obeys the differential equation

$$C\frac{dV}{dt} = -g_L(V - V_L) + g_L\Delta_T e^{(V-V_T)/\Delta_T} - I_{syn}(t) \tag{3}$$

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 $g_{syn}(t)$ changes via a learning procedure (see *Learning rule* subsection). 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 using a collection of oscillatory Poisson processes with rate ν and pairwise correlation c. Each of these cortical spike trains, which we refer to as daughters, is generated from a baseline oscillatory Poisson process $\{X(t_n)\}_n$, the mother train, which has intensity function $\lambda(1 + A\sin(2\pi\theta t))$ such that the spike probability at time point t_n is

$$P(X(t_n) = 1) \propto \int_{t_{n-1}}^{t_n} \lambda(1 + A\sin(2\pi\theta t))dt,$$

where A and θ are the amplitude and the frequency of the underlying oscillation, respectively; $t_{n+1} - t_n =: \delta t$ is the time step; and λ is the mother train rate. After the mother train is computed, each mother spike is transferred to each daughter with probability p, checked independently for each daughter. To fix the daughters' rates and the correlation between the daughter trains, the mother train's rate is given by $\lambda = \nu/(p * \delta t)$ where

$$p = \nu + c(1 - \nu). \tag{4}$$

In the STDP network (see ***Figure 1***, left) we consider two different mother trains to generate the cortical daughter spike trains for the two different MSN populations. 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 ***Kreitzer and Malenka*** (***2008***), 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 ***Fourcaud-Trocmé et al.*** (***2003***)). 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 $V_{syn} = 0\,mV$. 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 ***Equation 4***. 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 $p^D = 2/3\,p^I$. In most simulations, we set $A = 0$ to consider non-oscillatory cortical activity. We have also tested the learning rule when $A = 0.06$ and $\theta = 25\,Hz$ and obtained similar results.

The network has been integrated computationally by using the Runge-Kutta (4,5) method in Matlab (ode45) with 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 *Action and rewards* and *Example implementation* subsections), all populations stop receiving inputs from the cortex until all neurons in the network are in the resting state for at least $50\,ms$. During these silent periods, no MSN spikes occur and hence 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.

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. Specifically, there are variables g_{syn}, w for each corticostriatal synapse, A_{PRE} for each daughter train, 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

$$\frac{dg_{syn}}{dt} = \Sigma_j w(t_j)\delta(t - t_j) - g_{syn}/\tau_g, \tag{5}$$

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 (***Baladron et al., 2017***), which descends from earlier work (***Izhikevich, 2007***). The idea of this plasticity scheme is that an eligibility trace E (cf. ***Shindou et al.*** (***2018***)) 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 cell occurs, the associated variable increases by a fixed amount, and otherwise, it decays exponentially. That is,

$$\begin{aligned} \frac{dA_{PRE}}{dt} &= \left(\Delta_{PRE} X_{PRE}(t) - A_{PRE}(t)\right)/\tau_{PRE}, \\ \frac{dA_{POST}}{dt} &= \left(\Delta_{POST} X_{POST}(t) - A_{POST}(t)\right)/\tau_{POST}, \end{aligned} \tag{6}$$

where $X_{PRE}(t)$ and $X_{POST}(t)$ are functions set to 1 at times t when, respectively, a neuron that is pre-synaptic to the post-synaptic neuron, or the post-synaptic neuron itself, fires a spike, and are zero otherwise, while 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

$$\frac{dE}{dt} = \left(X_{POST}(t)\, A_{PRE}(t) - X_{PRE}(t)\, A_{POST}(t) - E\right)/\tau_E \tag{7}$$

implying that if a pre-synaptic neuron spikes and then its post-synaptic target follows, such that $A_{PRE} > 0$ and X_{POST} becomes 1, 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 1, then E decreases; at times without spikes, the eligibility decays exponentially with rate τ_E.

In contrast to previous work (***Baladron et al., 2017***), 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 (***Dreyer et al., 2010***; ***Richfield et al., 1989***; ***Gonon, 1997***; ***Keeler et al., 2014***) 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 changes 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

$$\frac{dw}{dt} = \alpha_w E f(K_{DA})(w^X_{max} - w), \tag{8}$$

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 ($X = D$) or an iMSN ($X = I$), c controls the saturation of weights to iMSNs, and $|\cdot|$ denotes the absolute value function. The dopamine level K_{DA} itself evolves as

$$\frac{dK_{DA}}{dt} = \Sigma_i (DA_{inc}(t_i) - K_{DA})\delta(t_i) - K_{DA}/\tau_{DOP}, \tag{9}$$

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:

$$\begin{aligned} DA_{inc}(t) &= r_i(t) - \max_i \left\{Q_i(t)\right\}, \\ Q_i(t+1) &= Q_i(t) + \alpha \left(r_i(t) - Q_i(t)\right), \end{aligned} \tag{10}$$

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 (***Eshel et al., 2015***), and $\alpha \in [0, 1]$ is the value learning rate. This rule for action value updates and dopamine release resembles past work (***Mikhael and Bogacz, 2016***) but uses a neurally tractable maximization operation (see ***Roesch et al.*** (***2007***); ***Kozlov and Gentner*** (***2013***) and references therein) to take into account that reward expectations may be measured relative to optimal past rewards obtained in similar scenarios (***Cohen et al., 2012***; ***Morris et al., 2006***). The evolution of these variables is illustrated in ***Figure 10***, which is discussed in more detail in the *Example implementation* subsection.

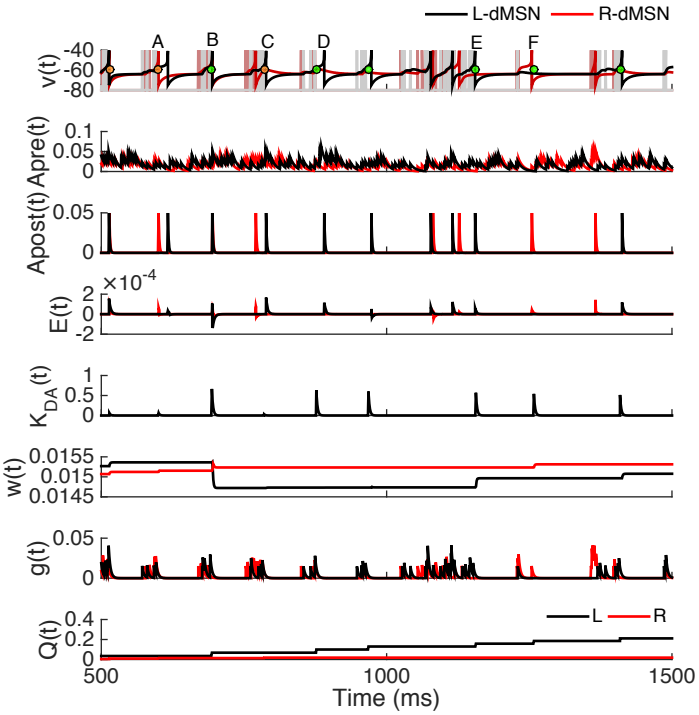


Figure 10. Evolution of the learning rule variables for particular dMSNs, one promoting the L action (black, actual reward value 0.7) and one promoting the R action (red, actual reward value 0.1). Each panel represents corresponding variables for both neurons except $K_{DA}(t)$, which is common across 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 the *Example implementation* subsection.

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}. When this condition occurs, a reward is delivered and the dopamine level is updated correspondingly, impacting all neurons in the network, depending on eligibility. 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. 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 (***Cui et al., 2013***), see *Results* section. 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 (data 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, to test the learning rule, 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 another case, we implement

probabilistic rewards: every time that an action occurs, the reward r_i is set to be 1 with probability p_i or 0 otherwise, $i \in \{L, R\}$. For this case, we consider three different probabilities such that $p_L + p_R = 1$ and $p_L > p_R$, keeping the action L as the preferred one. Specifically, we take $p_L = 0.85$, $p_L = 0.75$, and $p_L = 0.65$ to allow comparison with previous results (***Frank et al., 2015***). 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. Although the model gave sensible results, we did not explore this case thoroughly, and we simply show one example in ***Appendix 1***.

Example implementation

The algorithm for the learning rule simulations is as follows:

First, compute cortical mother spike trains and extract daughter trains to be used as inputs to each MSN 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 ***Equation 3*** and ***Equation 5***,
2. for each MSN, set the corresponding $X_{POST}(t)$ equal to 1 if a spike is performed or 0 otherwise and set the corresponding $X_{PRE}(t)$ to 1 if an input spike arrives or 0 otherwise,
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 spikes of the dMSNs in the ith population inside a time window consisting of the last $\Delta_{DA}\,ms$; if at least n_{act} spikes have occurred in this window, then action i has occurred and we update DA_{inc} and Q_i according to ***Equation 10***,
4. use RK45, with step size $dt = 0.01\,ms$, to solve ***Equation 6***-***Equation 8*** for each synapse, along with ***Equation 9***, yielding an update of DA and synaptic weight levels, for neurons that have $X_{PRE}(t) = 1$, update synaptic conductance using $g(t) = g(t) + w(t)$,
5. set $t = t + dt$.

Figure 10 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 ***Equation 7***. 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 ***Figure 10***. 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 action L.

Learning rule parameters

The learning rule parameters have been chosen to capture various experimental observations, including some differences between dMSN 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 (***Flores-Barrera et al., 2010***). Moreover, dMSNs spike more than iMSNs when both types receive similar cortical inputs (***Escande et al., 2016***). 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 (***Dreyer et al., 2010***; ***Richfield et al., 1989***; ***Gonon, 1997***; ***Keeler et al., 2014***). On the other hand, the baseline firing rates of iMSNs exceed the baseline of dMSNs (***Mallet et al., 2006***), 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 (***Baladron et al., 2017***), 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.

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 $\Delta_{PRE} = 0.1$; ***Baladron et al.*** (***2017***)), $\Delta_{POST} = 6$ (instead of $\Delta_{POST} = 0.006$; ***Baladron et al.*** (***2017***)), $\tau_E = 3$ (instead of $\tau_E = 150$; ***Baladron et al.*** (***2017***)), $\tau_{PRE} = 9$ (instead of $\tau_{PRE} = 3$; ***Baladron et al.*** (***2017***)), and $\tau_{POST} = 1.2$ (instead of $\tau_{POST} = 3$; ***Baladron et al.*** (***2017***)). Finally, $\alpha_w = \{80, -55\}$ (instead of $\alpha_w = \{12, -11\}$; ***Baladron et al.*** (***2017***)) and $w_{max} = \{0.1, 0.03\}$ (instead of $w_{max} = \{0.00045, 0\}$; ***Baladron et al.*** (***2017***)), 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.

CBGT network

The spiking CBGT network is adapted from previous work (***Wei et al., 2015***). Like the STDP model described above, the CBGT network simulation is designed to decide between two actions, a left or right choice, based on incoming sensory signals (***Figure 1***). The full CBGT network was comprised of six interconnected brain regions (see ***Table 3***), including populations of neurons in the cortex, striatum (STR), external segment of the globus pallidus (GPe), internal segment of the globus pallidus (GPi), subthalamic nucleus (STN), and thalamus. Because the goal of the full spiking network simulations was to probe the consequential effects of corticostriatal plasticity on the functional dynamics and emergent choice behavior of CBGT networks after learning has already occurred, CBGT simulations were conducted in the absence of any trial-to-trial plasticity, and did not include dopaminergic projections from the subtantia nigra pars compacta. Rather, corticostriatal weights were manipulated to capture the outcomes of STDP learning as simulated with the learning network (*STDP network* section) under three different probabilistic feedback schedules (see ***Table 4***), each maintained across all trials for that condition (N=2500 trials each).

Neural dynamics

To build on previous work on a two-alternative decision-making task with a similar CBGT network and to endow neurons in some BG populations with bursting capabilities, all neural units in the

CBGT network were simulated using the integrate-and-fire-or-burst model (***Smith et al., 2000***). Each neuron's membrane dynamics were determined by:

$$C\frac{dV}{dt} = -g_L(V - V_L) - g_T hH(V - V_h)(V - V_T) - I_{syn} \tag{11}$$

In ***Equation 11***, parameter values are $C = 0.5\,nF$, $g_L = 25\,nS$, $V_L = -70\,mV$, $V_h = -0.60\,mV$, and $V_T = 120\,mV$. When the membrane potential reaches a boundary V_b, it is reset to V_r. We take $V_b = -50\,mV$ and $V_r = -55\,mV$.

The middle term in the right hand side of ***Equation 11*** represents a depolarizing, low-threshold T-type calcium current that becomes available when h grows and when V is depolarized above a level V_h, since $H(V)$ is the Heaviside step function. For neurons in the cortex, striatum (both MSNs and FSIs), GPi, and thalamus, we set $g_T = 0$, thus reducing the dynamics to the simple leaky integrate-and-fire model. For bursting units in the GPe and STN, rebound burst firing is possible, with g_T set to $0.06\,nS$ for both nuclei. The inactivation variable, h, adapts over time, decaying when V is depolarized and rising when V is hyperpolarized according to the following equations:

$$\frac{dh}{dt} = \frac{-h}{\tau_h^-}, \text{when } V \geq V_h \tag{12}$$

and

$$\frac{dh}{dt} = \frac{1-h}{\tau_h^+}, \text{when } V < V_h \tag{13}$$

with $\tau_h^- = -20\,ms$ and $\tau_h^+ = 100\,ms$ for both GPe and STN.

For all units in the model, the synaptic current I_{syn}, reflects both the synaptic inputs from other explicitly modeled populations of neurons within the CBGT network, as well as additional background inputs from sources that are not explicitly included in the model. This current is computed using the equation

$$I_{syn} = g_1 s_1 (V - V_E) + \frac{g_2 s_2 (V - V_E)}{1 + e^{-0.062V/3.57}} + g_3 s_3 (V - V_I), \tag{14}$$

The reversal potentials are set to $V_E = 0\,mV$ and $V_I = -70\,mV$. The synaptic current components correspond to AMPA (g_1), NMDA (g_2), and $GABA_A$ (g_3) synapses. The gating variables s_i for AMPA and $GABA_A$ receptor-mediated currents satisfy:

$$\frac{ds_i}{dt} = \sum_j \delta(t - t_j) - \frac{s_i}{\tau} \tag{15}$$

while NMDA receptor-mediated current gating obeys:

$$\frac{ds_3}{dt} = \alpha(1 - s_3)\sum_j \delta(t - t_j) - \frac{s_3}{\tau} \tag{16}$$

In ***Equation 15*** and ***Equation 16***, t_j is the time of the j^{th} spike and α= 0.63. The decay constant, τ, was $2\,ms$ for AMPA, $5\,ms$ for GABA_A, and $100\,ms$ for NMDA-mediated currents. A time delay of $0.2\,ms$ was used for synaptic transmission.

Network architecture

The CBGT network includes six of the nodes shown in ***Figure 1***, excluding the dopaminergic projections from the substantia nigra pars compacta that are simulated in the STDP model. The membrane dynamics, projection probabilities, and synaptic weights of the network (see ***Table 3***) were adjusted to reflect empirical knowledge about local and distal connectivity associated with different populations, as well as resting and task-related firing patterns (***Wei et al., 2015***; ***Lo and Wang, 2006***).

The cortex included separate populations of neurons representing sensory information for L (N=270) and R (N=270) actions that approximate the processing in the intraparietal cortex or frontal

eye fields. On each trial, L and R cortical populations received excitatory inputs from an external source, sampled from a truncated normal distribution with a mean and standard deviation of $2.5\,Hz$ and 0.06, respectively, with lower and upper limits of $2.4\,Hz$ and $2.6\,Hz$. Critically, L and R cortical populations received the same strength of external stimulation on each trial to ensure that any observed behavioral effects across conditions were not the result of biased cortical input. Excitatory cortical neurons also formed lateral connections with other cortical neurons with a diffuse topology, or a non-zero probability of projecting to recipient neurons within and between action channels (see ***Table 3*** for details). The cortex also included a single population of inhibitory interneurons (CtxI; N=250 total) that formed reciprocal connections with left and right sensory populations. Along with external inputs, cortical populations received diffuse ascending excitatory inputs from the thalamus (Th; N=100 per input channel).

L and R cortical populations projected to dMSN (N=100/channel) and iMSN (N=100/channel) populations in the corresponding action channel; that is, cortical signals for a L action projected to dMSN and iMSN cells selective for L actions. Both cortical populations also targeted a generic population of FSI (N=100 total) providing widespread but asymmetric inhibition to MSNs, with stronger FSI-dMSN connections than FSI-iMSN connections (***Gittis et al., 2010***). Within each channel, dMSN and iMSN populations also formed recurrent and lateral inhibitory connections, with stronger inhibitory connections from iMSN to dMSN populations (***Gittis et al., 2010***). Striatal MSN populations also received channel-specific excitatory feedback from corresponding populations in the thalamus. Inhibitory efferent projections from the iMSNs terminated on populations of cells in the GPe, while the inhibitory efferent connections from the dMSNs projected directly to the GPi.

In addition to the descending inputs from the iMSNs, the GPe neurons (N=1000/channel) received excitatory inputs from the STN. GPe cells also formed recurrent, within channel inhibitory connections that supported stability of activity. Inhibitory efferents from the GPe terminated on corresponding populations in the the STN (i.e., long indirect pathway) and GPi (i.e., short indirect pathway). We did not include arkypalldal projections (i.e., feedback projections from GPe to the striatum; ***Mallet et al.*** (***2012***)) as it is not currently well understood how this pathway contributes to basic choice behavior.

Similar to the GPe, STN populations were composed of bursing neurons (N=1000/channel) with channel-specific inhibitory inputs from the GPe as well as excitatory inputs from cortex (the hyperdirect pathway). The since no cancellation signals were modeled in the experiments (see *Simulations of experimental scenarios* subsection), the hyperdirect pathway was simplified to background input to the STN. Unlike the striatal MSNs and the GPe, the STN did not feature recurrent connections. Excitatory feedback from the STN to the GPe was assumed to be sparse but channel-specific, whereas projections from the STN to the GPi were channel-generic and caused diffuse excitation in both L- and R-encoding populations.

Populations of cells in the GPi (N=100/channel) received inputs from three primary sources: channel-specific inhibitory afferents from dMSNs in the striatum (i.e., direct pathway) and the corresponding population in the GPe (i.e., short indirect pathway), as well as excitatory projections from the STN shared across channels (i.e., long indirect and hyperdirect pathways; see ***Table 3***). The GPi did not include recurrent feedback connections. All efferents from the GPi consisted of inhibitory projections to the motor thalamus. The efferent projections were segregated strictly into pathways for L and R actions.

Finally, L- and R-encoding populations in the thalamus were driven by two primary sources of input, integrating channel-specific inhibitory inputs from the GPi and diffuse (i.e., channel-spanning) excitatory inputs from cortex. Outputs from the thalamus delivered channel-specific excitatory feedback to corresponding dMSN and iMSN populations in the striatum as well as diffuse excitatory feedback to cortex.

Table 3. Synaptic efficacy (g) and probability (P) of connections between populations in the CBGT network, as well as postsynaptic receptor types (AMPA, NMDA, and GABA). The topology of each connection is labeled as either diffuse, to denote connections with a $P > 0$ of projecting to left and right action channels, or focal, to denote connections that were restricted to within each channel.

Connection	P	g (nS)	Topology	Receptor(s)
Ctx-Ctx	0.325	0.0127	diffuse	AMPA
Ctx-Ctx	0.325	0.15	diffuse	NMDA
Ctx-CtxI	0.181	0.013	diffuse	AMPA
Ctx-CtxI	0.181	0.125	diffuse	NMDA
Ctx-FSI	1.00	0.18	diffuse	AMPA
Ctx-dMSN	1.00	0.225	focal	NMDA, AMPA
Ctx-iMSN	1.00	0.225	focal	NMDA, AMPA
Ctx-Th	0.87	0.0335	diffuse	NMDA, AMPA
CtxI-CtxI	1.00	2.3125	diffuse	GABA
CtxI-Ctx	1.00	1.3125	diffuse	GABA
dMSN-dMSN	0.34	0.28	focal	GABA
dMSN-iMSN	0.34	0.28	focal	GABA
dMSN-GPi	1.00	1.44	focal	GABA
iMSN-iMSN	0.34	0.28	focal	GABA
iMSN-dMSN	0.38	0.28	focal	GABA
iMSN-GPe	1.00	3.05	focal	GABA
FSI-FSI	1.00	2.45	diffuse	GABA
FSI-dMSN	1.00	1.95	diffuse	GABA
FSI-iMSN	1.00	1.85	diffuse	GABA
GPe-GPe	0.05	1.50	diffuse	GABA
GPe-STN	0.05	0.40	focal	GABA
GPe-GPi	1.00	0.03	focal	GABA
STN-GPe	0.12	0.07	focal	AMPA
STN-GPe	0.12	4.00	focal	NMDA
STN-GPi	1.00	0.078	diffuse	NMDA
GPi-Th	1.00	0.142	focal	GABA
Th-Ctx	0.625	0.015	diffuse	NMDA
Th-CtxI	0.625	0.015	diffuse	NMDA
Th-dMSN	1.00	0.337	focal	AMPA
Th-iMSN	1.00	0.337	focal	AMPA
Th-FSI	0.625	0.30	diffuse	AMPA

Simulations of experimental scenarios

Because the STDP simulations did not reveal strong differences in Ctx-iMSN weights across reward conditions, only Ctx-dMSN weights were manipulated across conditions in the full CBGT network simulations. In all conditions the Ctx-dMSN weights were higher in the left (higher/optimal reward) than in the right (lower/suboptimal reward) action channel (see ***Table 4***). On each trial, external input was applied to L- and R-encoding cortical populations, each projecting to corresponding

populations of dMSNs and iMSNs in the striatum, as well as to a generic population of FSIs. Critically, all MSNs also received input from the thalamus, which was reciprocally connected with cortex. Due to the suppressive effects of FSI activity on MSNs, sustained input from both cortex and thalamus was required to raise the firing rates of striatal projection neurons to levels sufficient to produce an action output. Due to the convergence of dMSN and iMSN inputs in the GPi, and their opposing influence over BG output, co-activation of these populations within a single action channel served to delay action output until activity within the direct pathway sufficiently exceeded the opposing effects of the indirect pathway (***Wei et al., 2015***). The behavioral choice, as well as the time of that decision (i.e., the RT) were determined by a winner-take-all rule with the first action channel to cause the average firing rate of its thalamic population to rise above a threshold of $30\,Hz$ being selected.

Table 4. Corticostriatal weights in the CBGT network across levels of reward probability. Values of w were used to scale the synaptic efficacy of corticostriatal inputs ($g_{\text{Ctx-MSN}}$) to the direct (D) and indirect (I) pathways within the left (L) and right (R) action channels.

P(rew | Left)	$w_{D,L}$	$w_{I,L}$	$w_{D,R}$	$w_{I,R}$
Low	1.01	1.00	0.99	1.00
Med.	1.02	1.00	0.97	1.00
High	1.035	1.00	0.945	1.00

Drift Diffusion Model

To understand how altered corticostriatal weights influence decision-making behavior, we fit the simulated behavioral data from the CBGT network with a DDM (***Ratcliff et al.*** (***2016***); ***Ratcliff*** (***1978***)) and compared alternative models in which different parameters were allowed to vary across reward probability conditions. The DDM is an established model of simple two-alternative choice behavior, providing a parsimonious account of both the speed and accuracy of decision-making in humans and animal subjects across a wide variety of binary choice tasks (***Ratcliff et al., 2016***). It assumes that input is stochastically accumulated as the log-likelihood ratio of evidence for two alternative choices until reaching one of two decision thresholds, representing the criterion evidence for committing to a choice. Importantly, this accumulation-to-bound process affords predictions about the average accuracy, as well as the distribution of response times, under a given set of model parameters. The core parameters of the DDM include the rate of evidence accumulation, or drift rate (v), the distance between decision boundaries, also referred to as the threshold (a), the bias in the starting-point between boundaries for evidence accumulation (z), and a non-decision time parameter that determines when accumulation of evidence begins (tr), accounting for sensory and motor delays.

To narrow the subset of possible DDM models considered, DDM fits to the CBGT model behavior were conducted in three stages using a forward stepwise selection process. First, we compared models in which a single parameter in the DDM was free to vary across reward conditions. For these simulations all the DDM parameters were tested. Next, additional model fits were performed with the best-fitting model from the previous stage, but with the addition of a second free parameter. Finally, the two best fitting dual parameter models were submitted to a final round of fits in which trial-wise measures of striatal activity (see ***Figure 8***B-C) were included as regressors on the two designated parameters of the DDM. All CBGT regressors were normalized between values of 0 and 1. Each regression model included one regression coefficient capturing the linear effect of a given measure of neural activity on one of the free parameters (e.g., a, v, or z), as well as an intercept term for that parameter, resulting in a total of four free parameters per selected DDM parameter or 8 free parameters altogether. For example, in a model where drift rate is estimated as function of the difference between dMSN firing rates in the left and right action channels, the

drift rate on trial t is given by $v_j(t) = \beta_0^v + \beta_j^v \cdot X_j(t)$, where β_0^v is the drift rate intercept, β_j^v is the beta coefficient for reward condition j, and $X_j(t)$ is the observed difference in dMSN firing rates between action channels on trial t in condition j. A total of 24 separate regression models were fit, testing all possible combinations between the two best-fitting dual parameter models and the four measures of striatal activity summarized in ***Figure 8***B-C.

Fits of the DDM were performed using HDDM (see ***Wiecki et al.*** (***2013***) for details), an open source Python package for Bayesian estimation of DDM parameters. Each model was fit by drawing 2000 Markov Chain Monte-Carlo (MCMC) samples from the joint posterior probability distribution over all parameters, with acceptance based on the likelihood (see ***Navarro and Fuss*** (***2009***)) of the observed accuracy and RT data given each parameter set. A burn-in period of 1200 samples was implemented to ensure that model selection was not influenced by samples drawn prior to convergence. Sampling chains were also visually inspected for signs of convergence failure; however, parameters in all models showed normally distributed posterior distributions with little autocorrelation between samples suggesting that sampling parameters were sufficient for convergence. The prior distributions used to initialize all DDM parameters included in the fits can be found in ***Wiecki et al.*** (***2013***).

Acknowledgments

C. Vich 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, 1612913 (CRCNS), 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.

Competing Interests

The authors declare no financial or non-financial competing interests.

References

Afacan-Seref K, Steinemann NA, Blangero A, Kelly SP. Dynamic Interplay of Value and Sensory Information in High-Speed Decision Making. Current Biology. 2018; 28(5):795–802.

Alexander GE, Crutcher MD. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci. 1990 Jul; 13(7):266–271.

Baladron J, Nambu A, Hamker FH. The subthalamic nucleus-external globus pallidus loop biases exploratory decisions towards known alternatives: a neuro-computational study. European Journal of Neuroscience. 2017; p. 1–14. doi: 10.1111/ejn.13666.

Balleine BW, Delgado MR, Hikosaka O. The role of the dorsal striatum in reward and decision-making. J Neurosci. 2007 Aug; 27(31):8161–8165.

Baum CW, Veeravalli VV. A sequential procedure for multihypothesis testing. IEEE Transactions on Information Theory. 1994; 40(6).

Bogacz R. Optimal decision-making theories: linking neurobiology with behaviour. Trends in cognitive sciences. 2007; 11(3):118–125.

Bogacz R, Gurney K. The basal ganglia and cortex implement optimal decision making between alternative actions. Neural computation. 2007; 19(2):442–477.

Bogacz R, Larsen T. Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. Neural computation. 2011; 23(4):817–851.

Burnham KP, Anderson DR. Model Selection and Inference: A Practical Information-Theoretic Approach, vol. 80; 1998.

Caballero JA, Humphries MD, Gurney KN. A probabilistic, distributed, recursive mechanism for decision-making in the brain. PLoS Comput Biol. 2018 Apr; 14(4):e1006033.

Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. Nat Neurosci. 2008 Jun; 11(6):693–702.

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 Jan; 482(7383):85–88.

Collins AGE, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev. 2014 Jul; 121(3):337–366.

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 Feb; 494(7436):238–242.

Cui Y, Paillé V, Xu H, Genet S, Delord B, Fino E, Berry H, Venance L. Endocannabinoids mediate bidirectional striatal spike-timing-dependent plasticity. Journal of Physiology. 2015; 593(13):2833 – 2849. doi: 10.1113/JP270324.

Ding L, Gold JI. Caudate encodes multiple computations for perceptual decisions. J Neurosci. 2010 Nov; 30(47):15747–15759.

Donahue CH, Liu M, Kreitzer A. Distinct value encoding in striatal direct and indirect pathways during adaptive learning. bioRxiv. 2018; https://www.biorxiv.org/content/early/2018/03/07/277855, doi: 10.1101/277855.

Doya K. Modulators of decision making. Nat Neurosci. 2008 Apr; 11(4):410–416.

Draglia V, Tartakovsky AG, Veeravalli VV. Multihypothesis sequential probability ratio tests. I. Asymptotic optimality. IEEE Transactions on Information Theory. 1999; 45(7):2448–2461.

Dreyer JK, Herrik KF, Berg RW, Hounsgaard JD. Influence of Phasic and Tonic Dopamine Release on Receptor Activation. Journal of Neuroscience. 2010; 30(42):14273–14283. doi: 10.1523/JNEUROSCI.1894-10.2010.

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.

Dunovan K, Verstynen T. Believer-Skeptic meets Actor-Critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. Frontiers in neuroscience. 2016; 10:106.

Erlich JC, Brunton BW, Duan CA, Hanks TD, Brody CD. Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. Elife. 2015; 4:e05457.

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. Journal of Neuroscience. 2016; 36(21):5686–5698. doi: 10.1523/JNEUROSCI.0492-15.2016.

Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N. Arithmetic and local circuitry underlying dopamine prediction errors. Nature. 2015; 525(7568):243.

Flores-Barrera E, Vizcarra-Chacón B, Tapia D, Bargas J, Galarraga E. Different corticostriatal integration in spiny projection neurons from direct and indirect pathways. Frontiers in Systems Neuroscience. 2010; 4:15. doi: 10.3389/fnsys.2010.00015.

Fourcaud-Trocmé N, Hansel D, van Vreeswijk C, Brunel N. How spike generation mechanisms determine the neuronal response to fluctuating inputs. J Neurosci. 2003 Dec; 23(37):11628–11640.

Frank MJ. Linking Across Levels of Computation in Model-Based Cognitive Neuroscience. In: *An Introduction to Model-Based Cognitive Neuroscience* Springer, New York, NY; 2015.p. 159–177.

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 Jan; 35(2):485–494.

Gardner MPH, Conroy JS, Shaham MH, Styer CV, Schoenbaum G. Lateral Orbitofrontal Inactivation Dissociates Devaluation-Sensitive Behavior and Economic Choice. Neuron. 2017 Dec; 96(5):1192–1203.e4.

Gittis AH, Nelson AB, Thwin MT, Palop JJ, Kreitzer AC. Distinct roles of GABAergic interneurons in the regulation of striatal output pathways. J Neurosci. 2010; 30(6):2223–2234.

Gold JI, Shadlen MN. The neural basis of decision making. Annu Rev Neurosci. 2007 Jan; 30(30):535–561.

Gonon F. Prolonged and Extrasynaptic Excitatory Action of Dopamine Mediated by D1 Receptors in the Rat Striatum In Vivo. Journal of Neuroscience. 1997; 17(15):5972–5978.

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 Biology. 2015 01; 13(1):1–25. doi: 10.1371/journal.pbio.1002034.

Herz DM, Little S, Pedrosa DJ, Tinkhauser G, Cheeran B, Foltynie T, Bogacz R, Brown P. Mechanisms Underlying Decision-Making as Revealed by Deep-Brain Stimulation in Patients with Parkinson's Disease. Current Biology. 2018; 28(8):1169–1178.

Herz DM, Zavala BA, Bogacz R, Brown P. Neural correlates of decision thresholds in the human subthalamic nucleus. Current Biology. 2016; 26(7):916–920.

Izhikevich EM. Dynamical systems in neuroscience: the geometry of excitability and bursting. Computational Neuroscience, Cambridge, MA: MIT Press; 2007.

Jahfari S, Ridderinkhof KR, Collins AGE, Knapen T, Waldorp L, Frank MJ. Cross-task contributions of fronto-basal ganglia circuitry in response inhibition and conflict-induced slowing. bioRxiv. 2017; p. 199299.

Katz LN, Yates JL, Pillow JW, Huk AC. Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature. 2016 Jul; 535(7611):285–288.

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; 282:156–175.

Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. science. 2009; 324(5928):759–764.

Klaus A, Martins GJ, Paixao VB, Zhou P, Paninski L, Costa RM. The Spatiotemporal Organization of the Striatum Encodes Action Space. Neuron. 2017; 95(5):1171 – 1180.e7. doi: https://doi.org/10.1016/j.neuron.2017.08.015.

Kozlov AS, Gentner TQ. Central auditory neurons display flexible feature recombination functions. Journal of Neurophysiology. 2013; 111(6):1183–1189.

Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D. Neuroscience needs behavior: correcting a reductionist bias. Neuron. 2017; 93(3):480–490.

Kreitzer AC, Malenka RC. Striatal Plasticity and Basal Ganglia Circuit Function. Neuron. 2008; 60(4):543 – 554. doi: https://doi.org/10.1016/j.neuron.2008.11.005.

Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science. 2015 Jul; 349(6244):184–187.

Licata AM, Kaufman MT, Raposo D, Ryan MB, Sheppard JP, Churchland AK. Posterior Parietal Cortex Guides Visual Decisions in Rats. J Neurosci. 2017 May; 37(19):4954–4966.

Lo CC, Wang XJ. Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat Neurosci. 2006 Jun; 9(7):956–963.

Mallet N, Ballion B, Le Moine C, Gonon F. Cortical Inputs and GABA Interneurons Imbalance Projection Neurons in the Striatum of Parkinsonian Rats. Journal of Neuroscience. 2006; 26(14):3875–3884. doi: 10.1523/JNEUROSCI.4439-05.2006.

Mallet N, Micklem BR, Henny P, Brown MT, Williams C, Bolam JP, Nakamura KC, Magill PJ. Dichotomous Organization of the External Globus Pallidus. Neuron. 2012; 74(6):1075–1086.

Manohar SG, Chong TTJ, Apps MAJ, Batla A, Stamelou M, Jarman PR, Bhatia KP, Husain M. Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. Curr Biol. 2015 Jun; 25(13):1707–1716.

Marr D, Poggio T. From understanding computation to understanding neural circuitry. . 1976; .

Mikhael JG, Bogacz R. Learning Reward Uncertainty in the Basal Ganglia. PLoS Comput Biol. 2016 Sep; 12(9):e1005062.

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nature Neuroscience. 2006; 9:1057–1063.

Navarro DJ, Fuss IG. Fast and accurate calculations for first-passage times in Wiener diffusion models. J Math Psychol. 2009 Aug; 53(4):222–230.

Parker JG, Marshall JD, Ahanonu B, Wu YW, 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(7704):177.

Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. Psychonomic bulletin & review. 2017; 24(4):1234–1251.

Polanía R, Krajbich I, Grueschow M, Ruff CC. Neural Oscillations and Synchronization Differentially Support Evidence Accumulation in Perceptual and Value-Based Decision Making. Neuron. 2014; 82(3):709–720.

Ratcliff R. A theory of Memory Retrival. Psychol Rev. 1978; 85(2):59–108.

Ratcliff R, Frank MJ. Reinforcement-Based Decision Making in Corticostriatal Circuits: Mutual Constraints by Neurocomputational and Diffusion Models. Neural Comput. 2012; 24:1186–1229.

Ratcliff R, Smith PL, Brown SD, McKoon G. Diffusion Decision Model: Current Issues and History. Trends Cogn Sci. 2016 Apr; 20(4):260–281.

Rescorla RA, Wagner AR, et al. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory. 1972; 2:64–99.

Richfield EK, Penney JB, Young AB. Anatomical and affinity state comparisons between dopamine D1 and D2 receptors in the rat central nervous system. Neuroscience. 1989; 30(3):767 – 777. doi: https://doi.org/10.1016/0306-4522(89)90168-1.

Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nature Neuroscience. 2007; 10:1615–1624.

Schmidt R, Leventhal DK, Mallet N, Chen F, Berke JD. Canceling actions involves a race between basal ganglia pathways. Nat Neurosci. 2013 Aug; 16(8):1118–1124.

Schultz W, Apicella P, Scarnati E, Ljungberg T. Neuronal activity in monkey ventral striatum related to the expectation of reward. J Neurosci. 1992; 12(12):4595–4610.

Shadlen MN, Newsome WT. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of neurophysiology. 2001; 86(4):1916–1936.

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. 2018 Mar; .

Simen P, Cohen JD, Holmes P. Rapid decision threshold modulation by reward rate in a neural network. Neural Netw. 2006 Oct; 19(8):1013–1026.

Smith GD, Cox CL, Sherman SM, Rinzel J. Fourier analysis of sinusoidally driven thalamocortical relay neurons and a minimal integrate-and-fire-or-burst model. J Neurophysiol. 2000 Jan; 83(1):588–610.

Sutton RS, Barto AG, Book aB. Reinforcement Learning : An Introduction. . 1998; .

Tecuapetla F, Jin X, Lima SQ, Costa RM. Complementary contributions of striatal projection pathways to action initiation and execution. Cell. 2016; 166(3):703–715.

Tecuapetla F, Matias S, Dugue GP, Mainen ZF, Costa RM. Balanced activity in basal ganglia projection pathways is critical for contraversive movements. Nature communications. 2014; 5:4315.

Tort ABL, Komorowski RW, Manns JR, Kopell NJ, Eichenbaum H. Theta–gamma coupling increases during the learning of item–context associations. Proceedings of the National Academy of Sciences. 2009; 106(49):20942–20947. doi: 10.1073/pnas.0911331106.

Wei W, Rubin JE, Wang XJ. Role of the indirect pathway of the basal ganglia in perceptual decision making. J Neurosci. 2015; 35(9):4052–4064.

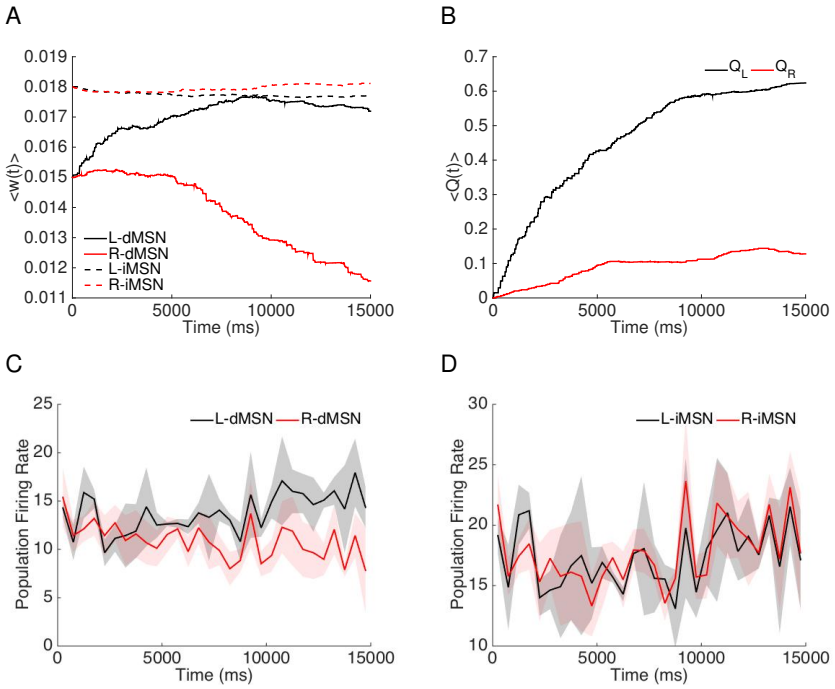
Wichmann T, DeLong MR. Functional and pathophysiological models of the basal ganglia. Current Opinion in Neurobiology. 1996; 6(6):751 – 758. doi: https://doi.org/10.1016/S0959-4388(96)80024-9.

Wiecki TV, Frank MJ. A computational model of inhibitory control in frontal cortex and basal ganglia. Psychol Rev. 2013 Apr; 120(2):329–355.

Wiecki TV, Sofer I, Frank MJ. HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. Frontiers in neuroinformatics. 2013; 7:14.

Yartsev MM, Hanks TD, Yoon AM, Brody CD. Causal contribution and dynamical encoding in the striatum during evidence accumulation. Elife. 2018 Aug; 7:e34929.

Supplementary Figures



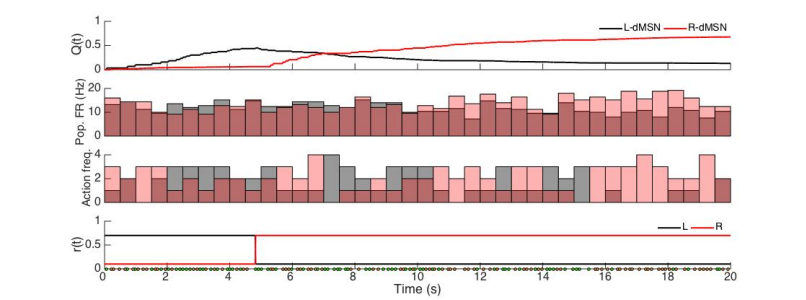
Supplementary Figure 1. Time courses of corticostriatal synapse weights and firing rates when the rewards are constant in time ($r_L(t) = 0.7$ and $r_R(t) = 0.1$). A: Averaged weights over 7 different realizations and over each of the four specific populations of neurons, which are dMSN selecting action 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.

Appendix 1

Results with step changes in action values

In ***Appendix 1 Figure 1*** we show the results of a simulation experiment with the STDP model in which the rewards associated with the L and R actions are switched after 5 sec. During the L-action consolidation period (from second 2 to 5), the firing rate for the L-dMSNs (D_L) becomes higher than that for the R-dMSNs (D_R). After $5\,s$, 20 L actions have been performened and the learning is almost consolidated, with $Q_L(t)$ and $Q_R(t)$ near $r_L = 0.7$ and $r_L = 0.1$ respectively (see first panel).

After the switch, there is a period of confusion where, even though L action is no longer the most rewarded, the network still shows a preference for L over R. Subsequently, the network learns that the R action is now more valuable than the L action, and the D_R grows while D_L decreases, such that eventually $D_R > D_L$. After 10.5 seconds or so, the rate of seleciton of R consistently that for L, showing the network's capacity for adjusting to reward changes.



Appendix 1 Figure 1. STDP results when the rewards associated with L and R actions are exchanged after learning is underway. 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.

Appendix 2

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 compute 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 $\Delta = 500\,ms$ as the number of occurrences of action i within that window divided by Δ.

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 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.