Received: 2 January 2020 | Revised: 23 March 2020 | Accepted: 25 March 2020

DOI: 10.1111/ejn.14745

SPECIAL ISSUE REVIEW

EJN European Journal of Neuroscience FENS WILEY

The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution

Jonathan E. Rubin[1] | Catalina Vich[2] | Matthew Clapp[3] | Kendra Noneman[4] | Timothy Verstynen[3,5]

[1]Department of Mathematics, Center for the Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department de Matemàtiques i Informàtica, Institute of Applied Computing and Community Code, Universitat de les Illes Balears, Palma, Spain

[3]Carnegie Mellon Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA

[4]Micron School of Materials Science and Engineering, Boise State University, Boise, ID, USA

[5]Department of Psychology, Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, USA

Correspondence
Jonathan E. Rubin, Department of Mathematics, 301 Thackeray Hall, University of Pittsburgh, Pittsburgh, PA, USA.
Email: jonrubin@pitt.edu

and
Timothy Verstynen, Department of Psychology, 342c Baker Hall, Carnegie, Mellon University, Pittsburgh, PA, USA.
Email: timothyv@andrew.cmu.edu

Funding information
European Regional Development Funds, Grant/Award Number: MTM2015-71509-C2-2-R, and MTM2017-83568-P; US National Science Foundation, Grant/Award Number: 1351748, 1516288, 1612913 and 1724240; Ministerio de Economía y Competitividad; Agencia Estatal de Investigación; Carnegie Mellon University

Abstract

The question of how cortico-basal ganglia-thalamic (CBGT) pathways use dopaminergic feedback signals to modify future decisions has challenged computational neuroscientists for decades. Reviewing the literature on computational representations of dopaminergic corticostriatal plasticity, we show how the field is converging on a normative, synaptic-level learning algorithm that elegantly captures both neurophysiological properties of CBGT circuits and behavioral dynamics during reinforcement learning. Unfortunately, the computational studies that have led to this normative algorithmic model have all relied on simplified circuits that use abstracted action-selection rules. As a result, the application of this corticostriatal plasticity algorithm to a full model of the CBGT pathways immediately fails because the spatiotemporal distance between integration (corticostriatal circuits), action selection (thalamocortical loops) and learning (nigrostriatal circuits) means that the network does not know which synapses should be reinforced to favor previously rewarding actions. We show how observations from neurophysiology, in particular the sustained activation of selected action representations, can provide a simple means of resolving this credit assignment problem in models of CBGT learning. Using a biologically realistic spiking model of the full CBGT circuit, we demonstrate how this solution can allow a network to learn to select optimal targets and to relearn action-outcome contingencies when the environment changes. This simple illustration highlights how the normative framework for corticostriatal plasticity can be expanded to capture macroscopic network dynamics during learning and decision-making.

Abbreviations: CBGT, cortico-basal ganglia-thalamic; DDM, drift-diffusion model.

Jonathan E. Rubin and Timothy Verstynen contributed equally to this article.

Edited by Dr. Yoland Smith.

The peer review history for this article is available at https://publons.com/publon/10.1111/ejn.14745

KEYWORDS
action selection, basal ganglia, credit assignment, dopamine, eligibility trace, reinforcement learning, synaptic plasticity

1 | INTRODUCTION

Survival in dynamic natural environments requires that animals effectively use prior experiences to guide future action selection. For example, remembering which plants previously provided a tasty treat versus which plants induced gastrointestinal distress can have a substantial impact on an animal's survival. In the mammalian brain, this form of reinforcement learning (Sutton, Barto, & Book, 1998) is driven in large part by the neurochemical dopamine (Schultz, 1998, 2016; Schultz & Romo, 1990). Sensory events that happen immediately after an action (e.g., a sweet taste or a painful sensation) can affect the output of midbrain dopamine neurons that target cells in the striatum, the major recipient of cortical inputs to the basal ganglia, tuning the sensitivity of striatal neurons to descending cortical inputs (for review, see Peak, Hart, & Balleine, 2019; Surmeier et al., 2010). In turn, this dopamine-dependent corticostriatal plasticity is thought to impact how the basal ganglia modulate cortical activity during action selection via their influence on thalamocortical connections, closing the so-called cortico-basal ganglia-thalamic (CBGT) computational "loop" (Alexander, DeLong, & Strick, 1986; DeLong & Wichmann, 2007; Parent & Hazrati, 1995).

The mainstream view of reward-related dopamine signals is that they shape corticostriatal synaptic plasticity in a way that promotes reward-inducing behaviors and suppresses non-rewarding actions (Mink, 1996), casting the corticostriatal synapses as a critical site for reward-dependent learning via synaptic plasticity. This elegant idea is consistent with a wealth of data across species (Lee, Tai, Zador, & Wilbrecht, 2015; Smeets, Marin, & Gonzalez, 2000). The challenge that has long stymied neuroscientists, however, is the proverbial devil in the details. The mapping from actions to dopamine release to the actual dopamine signals that are received in the striatum is complicated and depends on factors relating to many aspects of experimental conditions, including species involved, history preceding action performance, subject motivational state and measurement techniques. The transformation from dopamine signals to eventual modifications in behavior is even more opaque and difficult to elucidate because we lack a clear understanding of how the circuit-level properties of CBGT pathways map to behavior (but see Bogacz & Gurney, 2007; Dunovan, Vich, Clapp, Verstynen, & Rubin, 2019; Ratcli & Frank, 2012). This complexity provides a natural setting for the application of computational modeling, which can be used to instantiate the known components of the system together with other more exploratory ideas, both to test the capacity of these networks for effective learning and action selection, and to generate associated predictions.

Here, we review the computational approaches that have been used to understand how dopamine signals implement reinforcement learning in CBGT pathways. We structure the review part of this work to highlight the progression toward current efforts to connect synaptic-level models of plasticity with circuit-level models of CBGT-dependent decision-making and corresponding cognitive representations. We begin in Section 2 by providing a concise general description of the CBGT pathways, a brief introduction to ideas about evidence accumulation and an overview of the current understanding of corticostriatal plasticity. Next, in Section 3, we summarize previous computational models that describe how reward-related dopamine could influence corticostriatal synapses in a way that impacts future action selection. While these elegant synapse-level models are effective at capturing the nuanced dynamics of reinforcement learning, they rely on simple, abstracted action-selection rules that ignore or oversimplify the credit assignment problem (i.e., knowing how to change only those synapses that lead to the decision despite the spatiotemporal distance between selection and subsequent feedback signals). We go on to show how insights from the electrophysiology literature can provide essential clues as to how to resolve the credit assignment problem in CBGT pathways and present some proof-of-concept simulations to highlight how dopaminergic signals can implement reinforcement learning in a biologically plausible spiking model of CBGT circuits. Finally, we end by highlighting some open issues and future directions that the computational neuroscience literature can consider in order to bridge the gap between circuit properties of CBGT pathways and behavior.

2 | CIRCUIT-LEVEL ARCHITECTURE OF CBGT PATHWAYS

The computational goal of the CBGT loops appears to be to integrate information from competing cortical sources in order to (a) bias downstream selection systems and (b) use feedback signals to promote learning that modifies this biasing process in the future (Mink, 1996). The canonical model of how CBGT circuits (Figure 1) implement these computations relies on three structurally and functionally dissociable control pathways: the *direct* (facilitation), *indirect* (suppression) and *hyperdirect* (braking) pathways. This model assumes that the basal ganglia is organized into multiple action channels (Bogacz, 2007; Bogacz & Gurney, 2007; Mink, 1996), with each channel containing a direct and an indirect pathway. Activation of the direct pathway suppresses the basal ganglia output nuclei, here represented by the internal segment of the globus pallidus (GPi; the other major output nucleus being the substantia nigra pars reticulata). The standard theory of these pathways (Nambu, 2004) states that as the GPi tonically inhibits the thalamus, activation of the direct pathway relieves the thalamus from this inhibition, allowing it to facilitate action execution by activating or otherwise promoting specific action representations in cortical motor areas (Sauerbrie et al., 2020). This disinhibition of corticothalamic circuits is thought to be the basis for the selection of individual actions. In contrast, activation of the indirect pathway inhibits the external segment of the globus pallidus (GPe), which in turn inhibits both the subthalamic nucleus (STN) and the GPi. The net effect of indirect pathway activation is therefore enhancement of GPi activity, leading to greater suppression of the thalamus and, as a result, of cortical motor areas. Empirical evidence showing that, in freely moving mice, the expression of individual actions coincides with co-activation of spatially clustered populations of direct and indirect medium spiny neurons (MSNs) (Klaus et al., 2017) supports the assumption of this channel-like architecture of CBGT pathways.

Mounting evidence also motivates some adjustments to the standard theory. Specifically, recent experiments have established that both the direct pathway MSNs (dMSNs) and the indirect pathway MSNs (iMSNs) linked with a particular action are co-activated during action selection (Cui et al., 2013; Parker et al., 2018; Tecuapetla, Jin, Lima, & Costa, 2016; Tecuapetla, Matias, Dugue, Mainen, & Costa, 2014). This simultaneous activation of dMSN and iMSN populations challenges the traditional model of a strict isomorphism between dMSN and iMSN activity and excitation and inhibition of actions, respectively (Mink, 1996). These empirical observations, along with analysis of the topological organization of CBGT pathways, have led to more recent theoretical models proposing that, within an action channel, the dMSN and iMSNs work in a competitive manner to regulate the certainty of a given action decision (Bariselli, Fobbs, Creed, & Kravitz, 2018; Dunovan, Lynch, Molesworth, & Verstynen, 2015; Dunovan & Verstynen, 2016; Mikhael & Bogacz, 2016). This regulation may be circumvented by the hyperdirect pathway, which is activated by cortical input directly to the STN, yielding a strong, diffuse GPi activation. This signal can clear out any lingering activity from earlier processing before a subsequent action initiation, and it can act as a fast, reactive "brake" across action channels once progress toward action initiation is underway (Aron & Poldrack, 2006; Fife et al., 2017). It remains unclear whether hyperdirect pathways are specific to individual action channels or have a more distributed influence on action-selection processes.

A computational framework can provide a convenient abstraction from direct and indirect pathway neuronal activity into a precisely defined action-selection process. A useful way to think of decision-making is to envision the accumulation of evidence for or against available options. From a modeling viewpoint, abstracted away from details of neuronal implementation, this process can be represented either as a race among multiple evidence accumulators, each building toward a corresponding decision threshold, or as a competition among evidence streams, each vying to push a single evidence tracker toward a corresponding threshold. Both of these representations have been realized in many past works through the use of variants of a *drift-diffusion* model (DDM) in which evidence induces a drift, or directed movement, on an otherwise random walker in a spatial domain and decision thresholds are included as actual spatial boundaries of this domain (Ratcliff, 1978). Based on electrophysiological observations showing ramping activity in cortical sensory and motor planning areas, it was originally thought that the process of accumulation of evidence is implemented by cortical neurons (Gold & Shadlen, 2007), which pass their information on to striatal targets in the direct and indirect

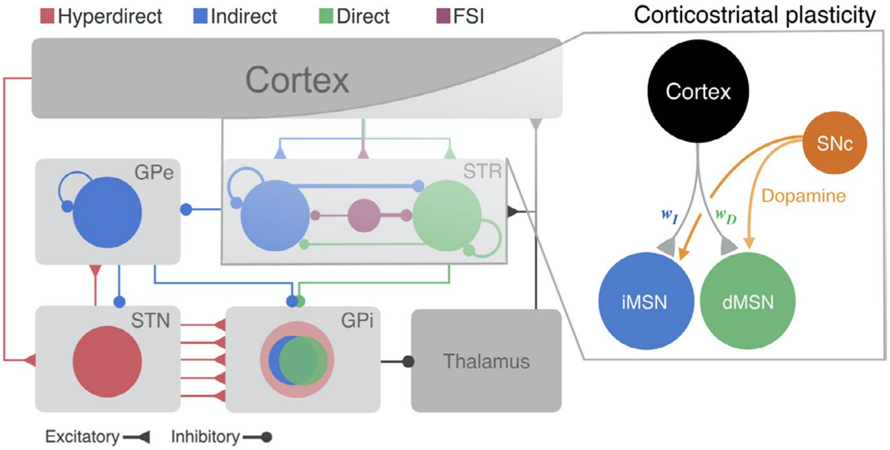


FIGURE 1 Circuit-level architecture of the cortico-basal ganglia-thalamic loop highlighting the major pathways within a single action channel. dMSN, direct pathway striatal neurons; FSI, fast spiking interneuron; GPe, globus pallidus external segment; GPi, globus pallidus internal segment; iMSN, indirect pathway striatal neurons; SNc, substantia nigra pars compacta; STN, subthalamic nucleus; STR, striatum; w_D, w_I, corticostriatal synaptic weights

pathways. However, this model of cortex as being critical for the process of evidence accumulation has recently come into question (Latimer, Yates, Meister, Huk, & Pillow, 2015), with some experimental evidence showing that disruption of cortical "accumulator" areas fails to disrupt the process of evidence integration (Katz, Yates, Pillow, & Huk, 2016; Li, Daie, Svoboda, & Druckmann, 2016). Populations of striatal neurons also show anticipatory ramping preceding action selection (Lauwereyns, Watanabe, Coe, & Hikosaka, 2002) and, in contrast to cortical perturbations, inactivation of the dorsolateral striatum compromises perceptual discrimination by mice, impacting the rate of information accumulation during decision processes (Yartsev, Hanks, Yoon, & Brody, 2018; however, see also Ding & Gold, 2013). Indeed, from a computational perspective, the architecture of CBGT pathways appears to be ideal for implementing an accumulation-to-bound style decision process (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz & Larsen, 2011; Dunovan & Verstynen, 2016).

Once evidence accumulation leads to an action, learning from the consequences of that action modifies the subsequent accumulation of evidence. Learning within CBGT pathways requires dopaminergic feedback from projections from the substantia nigra pars compacta (SNc) to corticostriatal synapses (Hollerman & Schultz, 1998; Perrin & Venance, 2019; Schultz, 1998; Figure 1, inset). Due to the opposing effects of dopamine on the direct and indirect pathways (Collins & Frank, 2014; Shan, Ge, Christie, & Balleine, 2014), these feedback signals are thought to reinforce rewarded actions while suppressing punished actions. The experimental literature has afforded several critical insights that have helped guide our understanding of the computational process of dopaminergic learning. First, there are phasic dopamine responses that correlate with post-action feedback signals in a way that resembles reward prediction errors (Schultz, Apicella, Scarnati, & Ljungberg, 1992; Schultz, Dayan, & Montague, 1997), such that the magnitude of evoked dopamine scales with properties of the received reinforcement signal relative to expectations (Fiorillo, Tobler, & Schultz, 2003; Tobler, Fiorillo, & Schultz, 2005). Second, the magnitude of dopamine signals determines the degree of effective plasticity at corticostriatal synapses, modulating their sensitivity to cortical afferents (Hernández-López et al., 2000; Thurley, Senn, & Luscher, 2008). Third, the nature of phasic dopamine-induced plasticity at corticostriatal synapses depends on the dopamine-receptor subtype involved and pre- and post-synaptic spike timing. The prevailing view (Shen, Flajolet, Greengard, & Surmeier, 2008) is that for D1-receptor-expressing MSNs, primarily dMSNs, higher dopamine promotes greater long-term potentiation (LTP) and lower dopamine promotes greater long-term depression (LTD). This effect is believed to be inverted in D2-receptor MSNs, primarily iMSNs, for which high dopamine promotes greater LTD and lower dopamine promotes greater LTP. It is important to note that this issue is not yet settled and that this relationship may vary across areas of the striatum and may depend on other factors such as striatal up and down states (Calabresi, Picconi, Tozzi, & Filippo, 2007; Pennartz, Ameerun, Groenewegen, & Lopes da Silva, 1993; Rusu & Pennartz, 2020; Thomas, Malenka, & Bonci, 2000). Finally, plasticity at corticostriatal synapses involves processes acting on multiple time scales. On short time scales (<100 ms), spike-timing-dependent plasticity (STDP) is observed (Fino, Glowinski, & Venance, 2005; Pawlak & Kerr, 2008), allowing for Hebbian learning. Whether or not STDP is expressed at corticostriatal synapses, however, is strictly determined by signals that happen at a much longer time scale (multiple seconds) (Fisher et al., 2017; Shindou, Shindou, Watanabe, & Wickens, 2019). Specifically, phasic dopamine causes potentiation only at corticostriatal synapses that were active in an appropriate time window some seconds before dopamine release, as indicated by some form of tag or marker. Theoretically, by allowing only the synapses that were active at the approximate time of action selection to be modified, this "eligibility trace'" (Houk, Adams, & Barto, 1995; Miller, 1988; Sutton & Barto, 1998) is believed to resolve the credit assignment problem by effectively allowing plasticity to take into account activity that occurred in the past.

While these various empirical observations serve as the basis for our current theoretical understanding of how action selection, dopaminergic learning and evidence accumulation are implemented in CBGT pathways, it is worth noting that theoretical models have also gone on to inform lines of empirical research. For example, the theoretical concepts of action channels (Bogacz, 2007; Bogacz & Gurney, 2007; Mink, 1996) and eligibility traces (Izhikevich, 2007b; Miller, 1988) preceded and motivated the experimental research that is now used to support their existence (e.g., Klaus et al., 2017 and Fisher et al., 2017; Shindou et al., 2019, respectively). Thus, the theoretical and empirical literatures on CBGT circuits have developed symbiotically over time, providing complementary insights into feedback learning and action selection.

3 | MODELS OF LEARNING & ACTION SELECTION IN CBGT NETWORKS

One of our goals in this review is to provide a critical evaluation of the current computational models of learning and action selection in CBGT networks. To understand where the field is now, we must first survey where we have been over the past several decades. We identify several key conceptual steps through which computational modeling progressed and the implications, or outcomes, that have arisen from those conceptual steps.

3.1 | Step 1: Early models of CBGT pathways

Many computational models have been developed over time to describe various aspects of action selection and reward-based learning (Buede, 2013). Here, we focus on network models of the basal ganglia that combine both learning and action selection, with the goal of identifying common algorithmic structures of learning as implemented in CBGT pathways.

Reinforcement learning in the basal ganglia, as mediated by dopamine, and temporal credit assignment have been a focus of computational models since the 1990s (Houk et al., 1995). Indeed, Houk and his co-authors (1995) initially proposed many critical concepts for how basal ganglia pathways, including dopaminergic inputs, could resolve the credit assignment problem that have only recently been validated by empirical work. Other early models of basal ganglia pathways that considered action-selection mechanisms did not include rewards based on chosen actions (see Beiser, Hua, & Houk, 1997; Gillies & Arbuthnott, 2000; Humphries, Stewart, & Gurney, 2006; Lo & Wang, 2006 and references therein). In some cases, dopamine was included as a parameter that could be hand tuned and affected the network's ability to make selections at all (Berns & Sejnowski, 1996, 1998; Gurney, Prescott, & Redgrave, 2001a, 2001b; Humphries et al., 2006; Sen-Bhattacharya et al., 2018), but learning dynamics were largely ignored. In contrast, the work of Frank (2005) is notable for combining hypotheses about action selection together with the mechanisms of reinforcement learning in basal ganglia pathways. This model separates striatal neurons into “Go” and “No-Go” populations, which map onto the D1/direct (selection) and D2/indirect (control) pathways, respectively. These pathways compete to control the activity downstream in the GPi, the level of which determines whether or not an action is selected. This model provided fundamental insights about the impacts of dopamine in the basal ganglia during reinforcement learning, suggesting that phasic dopamine modulates the basal ganglia to facilitate (Go population) or suppress (No-Go population) the response to a particular stimulus. In subsequent work, this modeling framework was extended to incorporate a role for the STN in producing a general braking signal, transmitted via a diffuse projection to the basal ganglia output layer (Frank, 2006). This function became especially important in high conflict scenarios (i.e., where the optimal decision is uncertain) featuring only subtle distinctions in the outcomes associated with different actions. In these conditions, suppression of premature decisions was found to be needed to achieve optimal reward outcomes. Learning in these models relied on both Hebbian plasticity (implemented via the Oja rule (Oja, 1982)) and error-driven temporal-difference learning (Sutton, 1988) to modify synaptic weights. Model neuronal activity was determined by evaluating sigmoidal activation functions, and the issue of eligibility was not considered.

3.2 | Step 2: Linking the CBGT network with cognitive representations of action selection

Building off of Frank (2006), Ratcliff and Frank (2012) provided more support for a role of the STN in modulating an effective decision threshold. They used a rate-based CBGT network to generate synthetic choice and response times and subsequently fit a DDM to the simulated behavioral data. This fitting allowed for a direct estimate of the decision threshold that they could relate to network parameters. In their fits, the height of the decision threshold tracked with the level of STN activity, suggesting that the indirect pathway plays a crucial role in modulating the amount of evidence needed before an action can be selected.

Bogacz and Gurney (2007) employed a similar computational approach to study the role of the STN in decision-making, albeit with the assumption that task proficiency had already been attained and hence without modeling of reinforcement learning. Using a firing rate model with an architecture developed previously (Gurney, Prescott, & Redgrave, 2001a, 2001b), these authors showed, similarly to Ratcliff and Frank (2012), that if STN activity represents a decision threshold related to the level of conflict or similarity in the salience of available options, then the cortical-basal ganglia circuit can implement the multihypothesis sequential probability ratio test (Draglia, Tartakovsky, & Veeravalli, 1999), which is the asymptotically optimal statistical test for decision-making. More specifically, this outcome emerges if the kth STN subpopulation expresses a firing rate $\exp(y_k(t))$ when given cortical input $y_k(t)$, and these subpopulation rates are summed and compressed logarithmically by local (i.e., GPe) inhibition to form a diffuse signal of size $\ln\left[\sum_{K=1}^{N}\exp\left(y_k(t)\right)\right]$ that is sent to downstream basal ganglia output neurons, the output of which forms the basis for action selection. Subsequently, Bogacz and Larsen (2011) incorporated a simple form of reinforcement learning based on reward prediction error into this firing-rate-based decision-making framework, albeit only for weights onto direct pathway striatal neurons. This relatively simple form of plasticity allowed the network to adaptively bias its selection behavior toward more strongly reinforced options in a manner resembling how mammals adapt their selections over time.

3.3 | Step 3: Dopamine-mediated plasticity, spike timing and synaptic eligibility

A key advance in the development of computational models in which feedback signals modulate action selection was the incorporation of more biologically detailed synaptic plasticity mechanisms involving phasic dopamine release. Hong and Hikosaka (2011) developed a firing rate model that implemented saccades to visual stimuli and included different rules for dopamine-mediated plasticity, involving different dopamine thresholds, in the direct and indirect pathways, viewed as action-promoting and action-suppressing, respectively. In each pathway, the corticostriatal weight changes required an eligibility signal, with a growth rate proportional to the product of pre- and post-synaptic activity levels, and the presence of acetylcholine, assumed to be released when a visual stimulus was present as well as at the start of each new block of task trials. The dopamine level in this model depends on the activity in the SNc and indicates the presence or absence of reward expectation; that is, dopamine is a prospective reward signal rather than a retrospective response to a recent prior reward. It is important to point out that the Hong and Hikosaka model (Hong & Hikosaka, 2011) and most of the other models discussed so far have relied on rate-based networks, which track firing rates but not individual spikes. In a spiking network model, it becomes possible to add further biological realism associated with corticostriatal plasticity. Specifically, in spiking models, corticostriatal synaptic weights can evolve through STDP (Cui et al., 2015; Fino et al., 2005; Fino & Venance, 2010; Shen et al., 2008), in which the relative timing of cortical and striatal spikes affects weight modifications. These models also allow for the inclusion of eligibility, such that only those striatal neurons that spike with appropriate timing relative to changes in dopamine levels are eligible for modification (Izhikevich, 2007a). Gurney, Humphries and Redgrave (2015) published a seminal spiking model of striatal populations that combines these elements. Without plasticity, action selection in their framework depends on the relative activity levels of D1 and D2 MSN subpopulations corresponding to a channel for that action, which jointly determine the intensity of the basal ganglia outputs that control selection. Changes in corticostriatal synaptic weights are proportional to the product of a neuron-specific eligibility term and a shared dopamine-dependent term. The former surges when an MSN fires and then decays, with a maximal amplitude determined by the timing of the MSN spike relative to that of its cortical input source. To generate the dopamine-dependent contribution, the authors computed functions that interpolated results of STDP experiments done with fixed dopamine levels, with distinct rules for cortical inputs to D1 versus D2 MSNs. With the incorporation of this timing- and dopamine-dependent plasticity, a model with spiking MSNs based on the Izhikevich framework (Humphries, Lepora, Wood, & Gurney, 2009; Izhikevich, 2002) and activity-based STN neurons, albeit lacking GPe neurons and the STN-GPe loop, successfully learned to select an action channel driven by stronger cortical input. It is worth noting that these results were obtained under the assumption of a fixed peak phasic dopamine level that decayed gradually over successive trials; scaling of dopamine levels based on reward predictions or value learning was not incorporated in these simulations.

More recently, Baladron, Nambu and Hamker (2019) adapted this spiking model into a more complete basal ganglia network. In their adaptation, corticostriatal synaptic weights obey a differential equation $dw/dt = \alpha E(t)\cdot\mathrm{DA}(t) - \delta X_{\mathrm{pre}}(t)$ where α denotes a learning rate, $E(t)$ is a time-dependent eligibility factor, $\mathrm{DA}(t)$ is a dopamine factor, δ is a decay rate, and $X_{\mathrm{pre}}(t)$ is zero except on time steps when the pre-synaptic neuron fires, in which case it is set to 1. The dopamine factor generally decays exponentially over time, except that when an action is rewarded, it jumps to a level $Re^{-N_c/60}$, where N_c denotes the number of consecutive previous rewarded trials, and when an action is not rewarded, it jumps to the level $-R$ for a parameter $R > 0$. The eligibility trace is updated each time that the pre- or the post-synaptic neuron spikes, such that it increases (decreases) when a pre (post)-synpatic neuron spike is followed by a post (pre)-synpatic neuron spike and then decays toward zero exponentially. Upper and lower bounds are imposed to constrain weight changes, and the sign of each weight change depends on whether the post-synaptic neuron is a dMSN or an iMSN. The model is tuned such that after a few trials of learning, cortical activity largely drives a single striatal dMSN population linked to the rewarded action, and thus, eligibility is limited to this dMSN population. The model also includes STN and GPe populations as part of the indirect pathway as well as basal ganglia outputs targeting thalamic neurons. Finally, thalamic spikes drive a linear integrator toward an action-selection threshold. This model shows that under changes in reward scenarios, activity in the STN-GPe loop can promote exploration and learning of alternative responses. This increases the flexibility of the simulated behavior compared with earlier spiking network models of basal ganglia pathways.

3.4 | Outcome 1: Theories about exploitation versus exploration

The results of the simulations by Baladron et al. (2019) raise an interesting issue that is relevant to CBGT-dependent learning: Reward maximization requires managing the trade-off between exploitation, or repeated selection of an action with a known outcome, and exploration, or variation in action choices that occurs to survey available reward opportunities (Baladron et al., 2019). Computational reasoning originally led to the proposal that certain forms of activity in the STN-GPe loop of the indirect pathway would promote exploration (Chakravarthy, Joseph, & Bapi, 2010), consistent with what Baladron et al. (2019) later found. This idea was first tested in a rate model (Kalva, Rengaswamy, Chakravarthy, & Gupte, 2012) and then in a spiking model, which yielded a link between dopamine-modulated STN-GPe synchrony and the tendency to explore available options (Mandali, Rengaswamy, Chakravarthy, & Moustafa, 2015). This relationship highlighted a potentially crucial role of the indirect pathway in the ability to modify the "greediness" (or bias toward exploitation) of action policies during learning. Yet, exploration may not be determined by indirect pathway efficacy alone. Humphries, Khamassi and Gurney (2012) used a rate model of basal ganglia pathways to evaluate the role of tonic dopamine in modulating the exploration–exploitation trade-off. Under their model, it is assumed that tonic dopamine levels adjust striatal responsiveness to cortical input, the tendency to explore increases as dopamine levels decrease, and the ratio of indirect to direct pathway striatal activity correspondingly grows.

Exploration becomes particularly important for making effective decisions under conditions with reward variability or risk. To consider risky scenarios, Balasubramani, Pragathi, Chakravarthy, Ravindran, and Moustafa (2014) implemented a risk function to describe a trade-off between the expected cumulative reward and the reward variance. According to their model, phasic dopamine signals tune action values over time and interact with serotonin pathways that scale both risk estimation and the time scale of reward prediction errors. This simplified model of reinforcement learning did not model network dynamics per se, but provided an intuitive mechanism for how dopamine and serotonin may work together to manage exploration under high variability or risk.

Later, Mikhael and Bogacz (2016) also considered both mean reward and reward variance more directly in the context of basal ganglia computations. This work was done using several highly reduced frameworks in which a "critic" component, representing action values, and an "actor" component, corresponding to tendency to choose an action, evolve over time based on reward prediction errors. In some of their simulations, they used the opponent actor learning (OpAL) model, in which a dual actor system, with distinct Go and No-Go components and tuning of their relative contributions by reward-related DA, replaces the standard single actor (Collins & Frank, 2014). Interestingly, the authors show that in this setting, the mean reward is encoded in terms of the difference in weights of D1 and D2 striatal neurons and the sum of these weights scales with the spread of rewards. Moreover, a prediction of this model is that high tonic DA levels promote the seeking of more risky options, in contrast to previous work (Humphries et al., 2012).

Finally, a follow-up to this work by Bogacz (2017), also in the OpAL framework, explores the idea that activity in the BG allows the downstream thalamic neurons to evaluate the utility of each available option. This evaluation is done by comparing expected positive outcomes or rewards, scaled by motivational state, with negative aspects such as energy expenditure or risk associated with an action. In this theory, motivation depends on dopaminergic activity D as $D/(1 - D)$, while positive and negative expectations are encoded in the synaptic weights of striatal Go and No-Go neurons, respectively. Importantly, however, no spike-timing aspects of plasticity or issues of synaptic eligibility are involved in this model.

3.5 | Outcome 2: Upward mapping with DA-mediated STDP

In the recent work, we have revisited the approach of linking from CBGT-based models to cognitive constructs relating to decision-making, now with the inclusion of dopaminergic plasticity and eligibility (Dunovan et al., 2019; Vich, Dunovan, Verstynen, & Rubin, 2020). This effort built from previous work (Baladron et al., 2019; Gurney et al., 2015; Mikhael & Bogacz, 2016) to model corticostriatal synaptic plasticity based on DA-mediated STDP (Vich et al., 2020). As a first step, we modeled only the dynamics of dMSN and iMSN populations and their cortical targets, using a simplified action-selection rule based on sequences of MSN spikes, with dMSNs and iMSNs promoting and blocking actions, respectively. In this model, each synaptic weight evolves according to the type of the MSN neuron involved in the synapse, taking into account that dMSN neurons are more responsive to phasic changes in dopamine, while iMSN neurons are largely saturated by tonic dopamine (Dreyer, Herrik, Berg, & Hounsgaard, 2010; Gonon, 1997; Keeler, Pretsell, & Robbins, 2014; Richfield, Penney, & Young, 1989). The differential equations governing weight changes are similar to that discussed in Section 3.3, again including an eligiblity term, but with different dopamine dependence for dMSNs versus iMSNs (see Appendix S1 for more details; Tables 1 and 2). When an action is performed, the level of dopamine increases to a value $\mathrm{DA}_{\mathrm{inc}}(t)$ determined by the reward prediction error, while eligibility is determined by an STDP rule, following the Baladron et al. (2019) model. Using this learning scheme, we show that plasticity driven by phasic dopamine yields rapid action value learning and selection of rewarded actions, robustly across reward

TABLE 1 Connection parameters in network model

Connection type	Connection probability	Synaptic conductance (nS)	Topology	Receptor(s)
Ctx-Ctx	0.13	0.0127	Diffuse	AMPA
Ctx-Ctx	0.13	0.15	Diffuse	NMDA
Ctx-CtxI	0.0725	0.013	Diffuse	AMPA
Ctx-CtxI	0.0725	0.125	Diffuse	NMDA
Ctx-FSI	0.45	0.132	Diffuse	AMPA
Ctx-d/iMSN	0.45	0.1286	Focal	AMPA
Ctx-d/iMSN	0.45	0.063	Focal	NMDA
Ctx-Th	0.35	0.03	Diffuse	AMPA, NMDA
CtxI-CtxI	1	1.075	Diffuse	GABA
CtxI-Ctx	0.5	1.05	Diffuse	GABA
dMSN-d/iMSN	0.135	0.28	Focal	GABA
dMSN-GPi	0.57	1.1	Focal	GABA
iMSN-iMSN	0.15	0.28	Focal	GABA
iMSN-dMSN	0.135	0.28	Focal	GABA
iMSN-GPe	0.74	1.65	Focal	GABA
FSI-FSI	0.85	1.15	Diffuse	GABA
FSI-dMSN	0.66	0.984	Diffuse	GABA
FSI-iMSN	0.62	0.984	Diffuse	GABA
GPe-GPe	0.02	1.5	Diffuse	GABA
GPe-STN	0.02	0.4	Focal	GABA
GPe-GPi	1	0.012	Focal	GABA
STN-GPe	0.0485	0.07	Focal	AMPA
STN-GPe	0.0485	4.01	Focal	NMDA
STN-GPi	1	0.0324	Diffuse	NMDA
GPi-Th	0.85	0.067	Focal	GABA
Th-Ctx	0.25	0.02	Diffuse	NMDA
Th-CtxI	0.25	0.015	Diffuse	NMDA
Th-d/iMSN	0.45	0.255	Focal	AMPA
Th-FSI	0.25	0.3	Diffuse	AMPA

scenarios. Most importantly, in this model the tendencies to select individual actions emerge largely through plasticity-driven tuning of the relative balance of direct and indirect pathway corticostriatal synapse weights within an action channel. A greater efficacy of dMSNs over iMSNs, where both the dMSNs and iMSNs are part of the same action channel, promotes the selection of that action. In contrast, increasing the relative efficacy of iMSNs within a channel decreases the likelihood of selection of the corresponding action.

To understand how these changes in corticostriatal synapses affect more realistic action-selection behavior, we incorporated the relative balance of dMSN and iMSN corticostriatal weights obtained from these STDP simulations into a full spiking CBGT network model (Dunovan

TABLE 2 Learning parameters in network model

Parameters	Value
τ_{DOP}	2
α	0.3
dMSN α_w	55
iMSN α_w	−45
dPRE	0.8
dPOST	0.04
τ_E	15
τ_{PRE}	15
τ_{POST}	6
w_{max}	0.2143
w_{init}	0.1286

et al., 2019). The accuracy levels and reaction times obtained from network simulations under different corticostriatal weight schemes were fit with a DDM to map upward from CBGT dynamics to cognitive decision-making parameters. Similar to both Ratcliff and Frank (2012) and Bogacz and Gurney (2007), we observed that the dynamics of the indirect pathway tuned the decision threshold. Specifically, the firing rates of all iMSNs across all action channels associated with the trial-by-trial variation in the boundary height. In contrast, the rate of information accumulation (i.e., drift rate in the DDM) for an action correlated with the relative asymmetry of dMSN-to-iMSN competition across action channels. Specifically, as the firing rates of the dMSNs increased over the rates of iMSNs in one channel, relative to the ratio in another channel, a larger drift rate was needed to fit network behavior. This work adds substantial nuance to our understanding of how feedback signals can change decision processes by showing that the difference in direct pathway activity between action channels controls the evidence accumulation or drift rate, while the total activity in the indirect pathway across channels controls the decision threshold.

3.6 | Outcome 3: Normative algorithm for corticostriatal synaptic plasticity

When we look across the spiking models proposed thus far, a general form of corticostriatal plasticity dynamics emerges. We summarize the normative form of this plasticity in Algorithm 1. The critical factors that contribute to this algorithm are the relative timing of cortical (pre-synaptic) and MSN (post-synaptic) spikes, the type of dopamine receptors on post-synaptic cells, the existence of the eligibility trace, and the direction and magnitude of the phasic dopamine response following action selection. Even though these four factors collectively describe local plasticity at corticostriatal synapses, this process requires careful integration across distributed parts of the CBGT circuit, ranging from descending cortical projections into MSNs to the selection processes in thalamocortical pathways to ascending dopamine signals from the SNc. The spatial and temporal separation of these processes represents a critical limitation in our understanding of corticostriatal plasticity: In a scenario with ongoing spiking activity, how do CBGT circuits know which populations to credit with a specific decision that contributed to a specific outcome?

Algorithm 1 Normative form of the plasticity algorithm for corticostriatal synapses. See Vich et al. (2020) for a specific implementation.

```
assume the existence of cortical spike train inputs to MSNs, grouped into different action channels;
while decision to select an action is not made do
    –synaptic eligibility decays;
    –synaptic conductance decays;
    if any MSN neuron spikes then
        –the cortical synapses onto that neuron become more eligible for subsequent plasticity;
        –each synapse's increase in eligibility depends on the time since the most recent spike of its
          presynaptic cortical neuron (shorter times yield larger increases);
        –neurotransmitter release leads to conductance increases in synaptic currents for its postsynaptic
          targets (e.g., neurons in GPe or GPi);
    end
    if any cortical neuron spikes then
        –neurotransmitter release leads to conductance increases in synaptic currents for its postsynaptic
          targets (e.g., dMSN and iMSN neurons);
        –the eligibility of its synapses onto MSNs decreases, with greater decreases for synapses onto MSNs
          that fired more recently;
    end
    –activity downstream from the striatum builds toward the threshold for the eventual selection of an
      action (for example, selection may require that BG outputs associated with a channel drop low enough
      to disinhibit downstream targets in thalamus and elsewhere);
    –various factors such as motivation, likely related to levels of tonic dopamine and other neuromodulators,
      affect the conditions needed for selection to occur;
end
if a reward occurs then
    –a phasic change in dopamine results, the amplitude of which increases with the difference between the
      reward value and the predicted reward value based on any action that preceded the reward;
    –the dopamine change is positive when rewards are underestimated and negative when they are
      overestimated;
    –a weight factor that multiplicatively scales the maximal conductance of each corticostriatal synapse is
      modified based on the dopamine level, the eligibility of that synapse, and whether the post-synaptic
      neuron is a dMSN (D1 receptors) or iMSN (D2 receptors);
end
```

A recent suggestion is that cholinergic signals (including pauses) could impact dopamine release in a way that introduces spatiotemporal modulations and that could support credit assignment (Zhang, Fisher, Oswald, Wickens, & Reynolds, 2019). While intriguing, this idea nevertheless leaves open questions of how cholinergic interneuron activity patterns are controlled and exactly how the resulting signals interact with dopamine and reward. In the next section, we suggest an alternative mechanism that could potentially resolve the credit assignment problem without requiring participation of additional striatal neuron populations.

4 | LEARNING IN A FULL SPIKING CBGT NETWORK

4.1 | The credit assignment problem

Our analysis of the existing computational neuroscience literature (Section 3) highlights a critical gap at both the theoretical and empirical levels. While biologically inspired spiking models have elegantly captured the dynamics of dopaminergic learning at corticostriatal synapses, these models have largely relied on either unrealistically sparse MSN firing or abstracted action-selection rules that compress the timing of the selection and feedback processes. As a result, they largely avoid the credit assignment problem.

Consider the simplified scenario shown in Figure 2 in which two dMSNs promote two different selection options (e.g., leftward, L-dMSN, or rightward, R-dMSN, movements). For simplicity, let us suppose that the selection of an action emerges from increased firing of the corresponding dMSN over a prolonged time period. In our example, the left action is the selected option. Some time after selection, a reward signal is delivered via the phasic dopamine response. Only those synapses that were active during the eligibility window (dashed box in Figure 2) are affected by the feedback signal, and the degree of the subsequent synaptic change depends on the overall activity of units in this eligibility window.

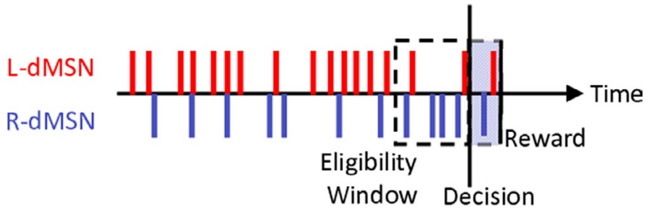


FIGURE 2 Schematic illustration of the credit assignment problem in a spiking neuronal network with the activity of distinct subpopulations of neurons representing evidence for two different options. The selection of one option (left movement) results from the relatively high firing rate of its subpopulation (L-dMSN) over a long time period, but that subpopulation need not fire more than the other in any constrained eligiblity time window preceding the decision point or reward delivery (shading denotes time period between the two). The units of time here are not specified because the possible effect being illustrated depends only on the duration of the eligibility window relative to the full decision process, not on absolute durations. Also, although only dMSNs are shown, iMSNs are also active during the same time period

While the overall firing rate, across time, of the L-dMSN is much higher than that of the R-dMSN, this activity is not uniform. In this example, the spike that eventually triggers the decision happens to occur after a brief window of relatively sparse L-dMSN firing. Because both dMSNs are active during the eligibility window, and in this case, the R-dMSN fires slightly more, *both* channels will be rewarded, with greater reward accruing to the right channel, even though that channel was not selected. Moreover, at the level of the striatum, the network has no way to know when a decision has been made. Thus, there is no reason to exclude the spikes that occur in between the decision time and the time of reward delivery (shaded blue part of the eligibility window in Figure 2), even though they did not contribute to the selection of the implemented action. Thus, in this simple model, the network has trouble giving credit to the dMSN that actually contributed to the downstream decision.

A more realistic scenario would feature action selection by populations of neurons, rather than individual cells, with progressive changes in synaptic weights over multiple repetitions of evidence gathering, action selection and reward-related dopamine release. In theory, averaging over these multiple trials and neurons could bias credit assignment in favor of the appropriate dMSNs, even if the wrong ones get credit from time to time. In our simulations, this theoretical possibility does not necessarily materialize. For example, in some instances, problematic credit assignment on early trials locks in sub-optimal selections, before superior options can be explored. Moreover, learning new behaviors after changes in action-outcome associations remains problematic, as a high level of activity in at least some of the neurons associated with previously rewarded actions occurs when the newly rewarded actions are selected. This erroneous credit assignment preserves behaviors that have become sub-optimal and prevents an efficient transition to newly advantageous behaviors.

This credit assignment problem is exacerbated by several factors in real CBGT pathways. First, both dMSNs and iMSNs fire during the deliberation and execution of actions (Cui et al., 2013; Tecuapetla et al., 2014, 2016). Indeed, if we suppose that R-iMSN activity is enhanced leading up to selection of a left action that elicits a reward, then eligibility of R-iMSNs will cause their synapses with cortex to be weakened, which will counterproductively interfere with selection of the left action on subsequent trials.

Second, the selection process happens much farther downstream in the thalamocortical pathways, often separated by hundreds of milliseconds or more from the dynamics in the striatum that contribute to the decision (see Figure 1). Finally, the duration of actual eligibility windows has not been pinned down precisely but seems to include spikes occurring approximately 1 s, but not 2 s, before a positively conditioned cue (Fisher et al., 2017; Yagishita et al., 2014) or spikes occurring about 2 s, but not 4 s, before dopamine release (Shindou et al., 2019), both of which include spikes that are significantly separated in time from reward signals.

Put all together, the structure of current STDP models of corticostriatal plasticity, the complexity of the selection process and the long time window between selection dynamics in the striatum and feedback signals all conspire to interfere with any simple translation of dopaminergic STDP models into spiking models of the entire CBGT circuit during natural decisions.

4.2 | Solving the credit assignment problem with sustained activity of selected channels

The structure of CBGT pathways, and in particular the distance between the dynamics of corticostriatal pathways that effectively drive the integration process and the thalamocortical pathways that eventually trigger an action, suggests two possible options for resolving the credit assignment problem: fundamentally rethinking the structure of the learning algorithm itself or rethinking the network dynamics involved in implementing the algorithm. The problem with the first option is that the algorithmic descriptions of STDP and dopaminergic learning have proven to be highly effective at both explaining and, in some cases, predicting empirical findings (see Section 3). Adding further complexity to these models risks reducing their extensibility.

So that leaves the option of rethinking the dynamics of network activity relevant for the action-selection process. Here, we can turn to the electrophysiological literature for some inspiration. In particular, it is useful to consider the seminal study by Cisek and Kalaska (2005) on movement representations across selection and execution stages of action decisions. In this study, the authors recorded from dorsal premotor (PMd) cortical neurons while monkeys deliberated and executed reaches to one of two spatially cued targets (Figure 3). At the beginning of each trial, a monkey was presented with a cue indicating two possible reach targets. Shortly after cue onset, the authors observed sustained activity in units that were directionally tuned to each of the corresponding movement directions. After a brief period of time, a new selection cue was presented indicating which of the two targets the monkey should reach for. Importantly, the monkey was trained not to make the reach upon perception of the selection cue, but to hold until a release cue was presented. Shortly after the selection cue appeared, the cortical units representing the unselected action reduced their firing, while the units representing the cued movement direction maintained sustained and even amplified activity until a short time after the release cue was delivered to signal that it was time to reach. This recruitment of multiple action representations and maintenance of the selected action representation have been replicated in other cortical areas (Coallier, Michelet, & Kalaska, 2015; Klaes, Westendorff, Chakrabarti, & Gail, 2011; Pastor-Bernier & Cisek, 2011) and are supported by psychophysical experiments as well (Gallivan, Logan, Wolpert, & Flanagan, 2016; Gallivan, Stewart, Baugh, Wolpert, & Flanagan, 2017; McKinstry, Dale, & Spivey, 2008, though see also Dekleva, Kording, & Miller, 2018).

Although these findings were obtained in cortical motor planning neurons, their dynamics across the deliberation and maintenance stages of action selection provide a critical insight into how broader CBGT action channels (which include PMd cortical units) may function. Specifically, if representations of the selected action channel are maintained until or somewhat beyond movement onset, while unselected

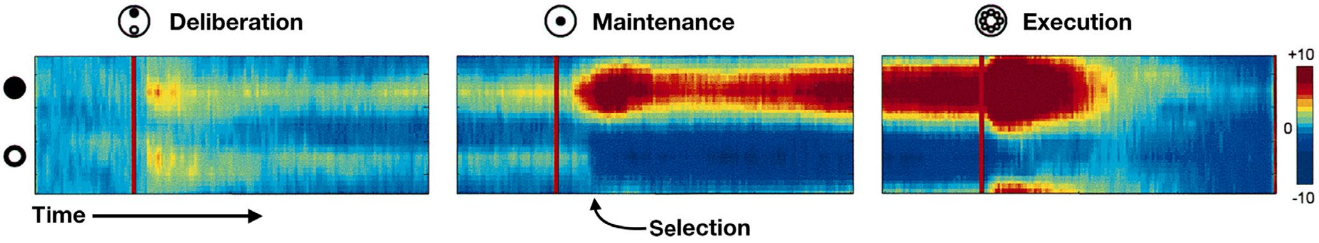


FIGURE 3 Normalized firing rates of cells in the macaque PMd that are tuned to one of two visual targets of reaching movements. At the onset of the trial, the monkey is presented with a cue indicating that either target may be selected (Deliberation stage). During this window, units representing both targets are active. After a predetermined time, the monkey is cued as to which target should be selected, but not given an imperative signal to begin moving (Maintenance stage). After the arrival of the selection cue, the units for the selected action continue to fire at an increased rate, while units for the unselected action cease firing after a short time. The units representing the selected action continue firing until some time after the imperative cue (indicating the start of the Execution stage) is delivered. Figure reprinted with permission from Cisek and Kalaska (2005)

channels return to baseline rates, then the eligibility for plasticity when dopaminergic feedback signals arrive will be much stronger in selected channels than in unselected channels. Thus, sustained activation of selected action channels would effectively resolve the credit assignment problem in CBGT networks without reconfiguring the general form of corticostriatal plasticity (Algorithm 1).

To test this hypothesis, we modified a previously published spiking network model of CBGT networks (Dunovan et al., 2019) (Figure 1) to have two novel properties: (a) a 300-ms delay between selection at the thalamus (first unit to reach >30 Hz firing) and the dopamine response, designed to mimic the delay between selection, execution and sensory signals of post-action feedback, and (b) a sustained activation of only those cortical populations representing the selected channel during the delay. Note that these populations project both to dMSNs and to iMSNs in that channel. Learning was implemented at the corticostriatal synapses to all MSN populations using the normative model of plasticity shown in Algorithm 1 (using the instantiation reported in Dunovan et al. (2019) and Vich et al. (2020)). The model is described in general terms in Appendix S1. All simulation code can be found at https://github.com/CoAxLab/AdaptiveCBGT.

Figure 4 shows the example network dynamics from a set of simulated trials where the network chooses between moving left (blue) and moving right (red). In this example simulation, left actions are always rewarded, while right actions are not. In all simulations, the sensory signals for left and right actions have the same signal-to-noise ratio; thus, any changes in behavior should be exclusively driven by feedback learning.

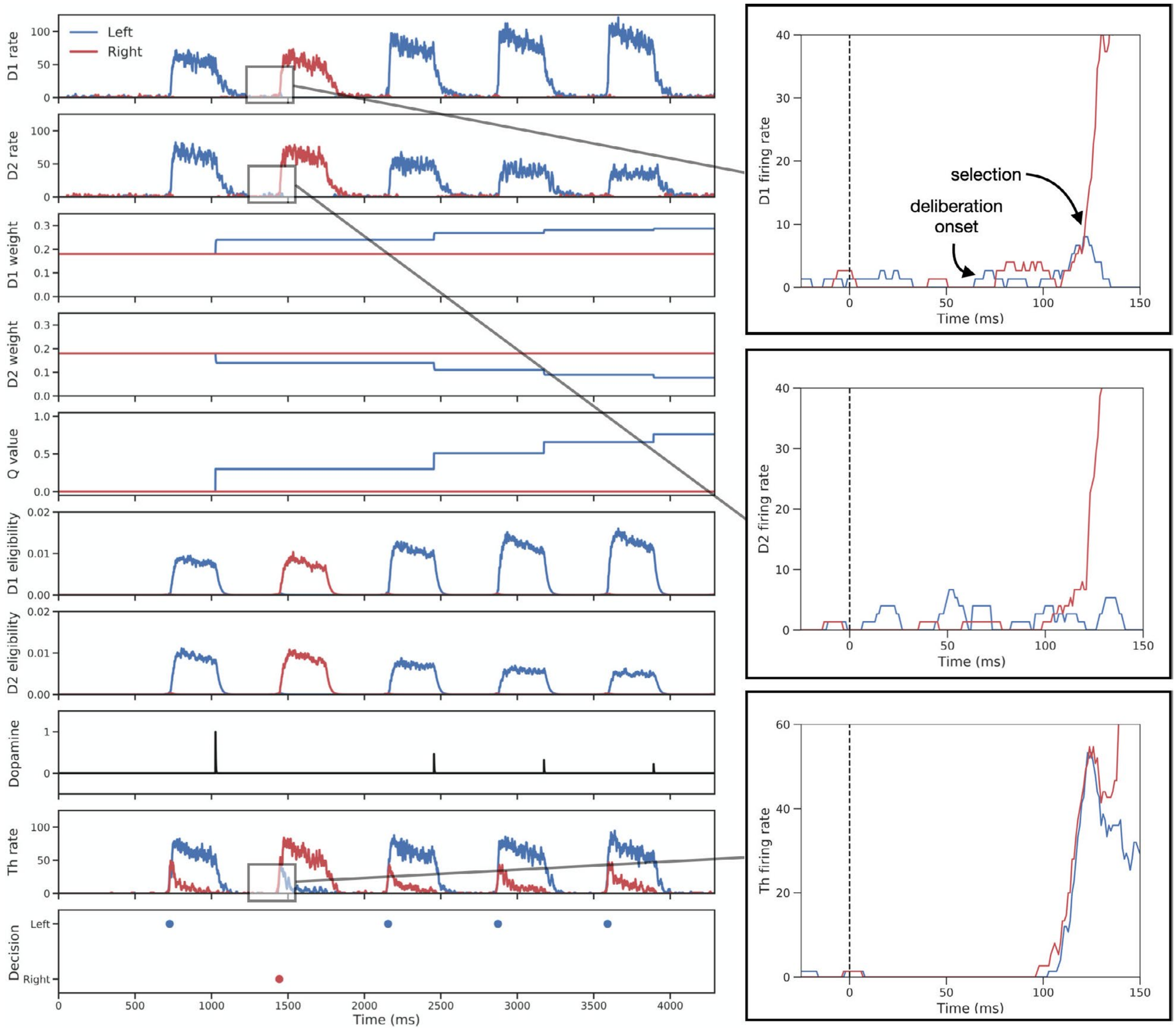


FIGURE 4 An example of five simulated trials in a deterministic reward task (see main text). Only left actions are rewarded. Inset images show close up of dMSNs (D1, top), iMSNs (D2, middle) and thalamic units (bottom) for a single trial. D1, dMSNs; D2, iMSNs; *Q* value, value associated with each action; Th, Thalamus

At the beginning of the example simulation, the network makes one left action (correct) and then one right action (incorrect). The upper inset shows the increase in dMSN firing rates between cue onset and selection (~125 ms), which is determined by activity downstream in the thalamus (lower inset), on the second trial. As the incorrect direction was chosen on that trial, there is no dopamine release and no corticostriatal synaptic weight changes. More generally, notice that when a dopamine burst appears in Figure 4, the corticostriatal synaptic weights (D1 weight, D2 weight) change only in the selected channel. This outcome indicates that the network has resolved the credit assignment problem. Based on the selection pattern over five trials (bottom row; Figure 4), it appears that the network quickly learns to select the appropriate action.

To test the limits and flexibility of this learning in the full spiking CBGT network model, we ran a set of 250 simulations where initially the left action was always rewarded, and then, after the 20th trial, the outcome contingencies switched, such that the right action was rewarded while the left action ceased to be rewarded. This switching experiment allows us to evaluate not only the effective learning of the network but also its flexibility. Figure 5 shows the trialwise selection probabilities across all simulated runs, in cases with (blue) and without (orange) sustained cortical activity after selection. With the sustained activity, during the initial learning phase, the network quickly stabilizes to selecting the left action most of the time, asymptoting after 8–10 trials. The reason that the selection probabilities do not asymptote closer to 100% has to do with firing rate variability at the thalamus, which has a nonzero intrinsic spike rate, and the intentional choice of a parameter scheme that does not allow the corticostriatal weights to diverge too far apart, which allows for flexibility. Indeed, in this parameter regime, after the outcome contingencies are switched (red dashed line), the network quickly relearns to prefer the right movement, stabilizing again after about 10–12 more trials. In contrast, without sustained activation, learning does not occur for these parameter values (orange trace; Figure 5). We tried multiple parameter schemes with this baseline network. For certain parameter schemes, partial learning could occur during the initial 20 trials; however, the network could not adapt to the switching of outcomes (data not shown).

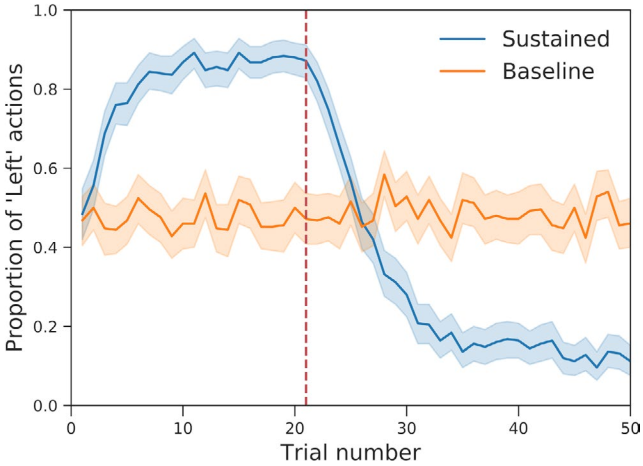


FIGURE 5 Mean selection probability on each trial across 250 simulated runs of the adaptive network, with (blue) and without (orange) sustained cortical activity after selection. In both cases, initially only "left" actions are rewarded; after 20 selections, the outcomes are switched and only "right" actions are rewarded. The red dashed line marks the trial when reward contingencies are switched. The cloud around the mean line shows the trialwise 95% confidence intervals. Note that learning occurs only when the sustained activity is present

4.3 | Lessons learned and some implications

These proof-of-concept simulations highlight how the credit assignment problem during reinforcement learning can be easily resolved by adding a single, neurophysiologically supported assumption: sustained activation of selected action channels, originating in cortex and propagating to the corresponding dMSN and iMSN populations. This phenomenon could manifest from many possible sources, including attentional mechanisms, working memory processes, or the overlap of selection, planning and execution representations in motor networks that would be engaged until the action is completed. We take no hard stance on the origins of this sustained activation. Instead, we only highlight it as a simple potential mechanism for how normative STDP processes at the corticostriatal synapses (Algorithm 1) can implement effective learning in a spiking network with realistic striatal spiking dynamics.

In our past work, we observed that dopamine-dependent corticostriatal plasticity, based on synaptic eligibility and relative cortical and striatal spike timing, leads to increased activation of both dMSN and iMSN populations in the same channel, because of the commonality of their inputs (Dunovan et al., 2019). This co-activation has been observed experimentally and runs counter to classic theories of the basal ganglia's role in action selection (Gurney, Prescott, & Redgrave, 2001a; Kropotov & Etlinger, 1999; Mink, 1996). If iMSNs fire before a consistently selected and rewarded action, then the cortical synapses to those iMSNs should weaken. This means, however, that co-activation might be expected to diminish over time. It is not clear why such a reduction in activity has not been observed. Sustained cortical activation within a rewarded channel could be strong enough to maintain this co-activation even in the face of weakened cortico-iMSN synapses, and hence, this idea offers a potential explanation for yet another perplexing experimental finding.

5 | FUTURE DIRECTIONS

Our review of the computational neuroscience literature shows an emerging general view of how dopamine-dependent plasticity of corticostriatal synapses can alter the dynamics and output of CBGT circuits in a way that promotes the selection of rewarding actions (Algorithm 1). While the assumptions underlying these models are derived from experimental observations, models necessarily represent simplifications of reality. A major issue that existing models have failed to address in a biologically plausible way is the credit assignment problem. That is, in a winner-take-all or sparse firing regime, only those neurons that drive a selected action are active before it occurs, and there is no difficulty in ensuring that reward signals only strengthen the synapses to these driver neurons. But biological details complicate this picture. In a setting in which decision-making follows after a temporally extended window of neuronal activity, during which neurons promoting as well as neurons inhibiting multiple actions are all active at rates that may gradually increase, credit assignment becomes messy and problematic. In Section 4, we propose and demonstrate computationally how the maintenance of activity associated with an action from selection through execution and feedback, observed cortically (Cisek & Kalaska, 2005), can provide a simple solution to the credit assignment problem.

The idea that competing action representations are activated during motor planning under uncertainty and that the selected representation is maintained after the decision is implemented is supported by both electrophysiological (Cisek & Kalaska, 2005; Coallier et al., 2015; Klaes et al., 2011; Pastor-Bernier & Cisek, 2011) and psychophysical data (Gallivan et al., 2016, 2017; McKinstry et al., 2008). Indeed, these findings have served as the basis for a growing computational modeling literature on cortical motor planning (Christopoulos, Bonaiuto, & Andersen, 2015; Cisek, 2007). It is worth pointing out that the simultaneous activation of multiple action representations prior to selection is not always observed experimentally. For example, Dekleva et al. (2018) found that when fitting PMd neuron activity during a multi-choice task to a single-trial analysis model, the best fit model suggested the presence of only a single motor plan during deliberation (Dekleva et al., 2018). While the authors argue that this discrepancy is related to the difference between single-trial analysis and trial averaging, the differences in the results obtained by Dekleva et al. (2018) from those of previous studies could also be due to fundamental differences in task design, which may have promoted a single-target selection strategy that reduced uncertainty as to the upcoming action choice. Importantly for the purposes of our model as presented here, the activity in these PMd neurons (a primary motor planning area) was sustained throughout the deliberation, selection and execution stages of the trial.

While most of the evidence we used to justify sustained activation of selected actions comes from recordings in cortical motor areas, it is not unexpected that such activation would propagate throughout the basal ganglia and thalamic nuclei connected to these cortical regions. For example, in settings where a cue signals an expectation of a reward, activity in thalamus is known to keep ramping from cue until reward delivery (Komura et al., 2001). It is therefore reasonable to conjecture that there may be similar ramping in thalamic neurons that project to the striatum and are associated with an action channel, persisting from the selection of that action until reward delivery and dopamine release. Presumably, output from thalamic areas associated with the selected action to their targets in the striatum could therefore serve as an alternative source of a positive feedback signal to achieve credit assignment. Of course, full representation of credit assignment would also require characterizing the microscale processes that implement the tagging of synapses, corresponding to the eligibility trace $E(t)$ in our model (Vich et al., 2020), but this aspect of the process is outside the scope of our current efforts and of the other models that we have reviewed.

It is also worth noting that the dynamics of dopamine in the proof-of-concept model presented here, as well the preceding modeling approaches that inspired it, remains rather simple and strongly connected to classic results on reward prediction error (Schultz, 1998; Schultz et al., 1992; Schultz & Romo, 1990). In reality, the dynamics of dopamine and its relation to feedback signals for learning are much more complicated. First, the dopamine signal comprises several components that may include a tonic level as well as multiple distinct phasic release events. Tonic dopamine may be a passive signal that must exceed a threshold to allow movement or may be linked to motivation, vigor and satiety (Hamid et al., 2016; Niv, Daw, Joel, & Dayan, 2007). Previous computational reinforcement learning models have included tonic dopamine by allowing it to control the learning rate (Beeler, Daw, Frazier, & Zhuang, 2010), MSN excitability (Gurney et al., 2015; Humphries et al., 2012), or STN-GPe connectivity (Chakravarthy et al., 2010; Mandali et al., 2015), each of which is assumed by the authors to affect the exploration–exploitation trade-off. But these models do not include dynamic mechanisms for modulating tonic dopamine levels, instead treating it as a parameter that is tuned by hand. In addition to a post-reward component, phasic dopamine release may include an early generalized response associated with the expectation of any reward (Nomoto, Schultz, Watanabe, & Sakagami, 2010); a response, following a specific stimulus that signals an imminent reward, which scales with expected reward size (Alves da Silva, Tecuapetla, Paixão, & Costa, 2018; Cohen, Haesler, Vong, Lowell, & Uchida, 2012; Nomoto et al., 2010); or a response that occurs after movement but before reward delivery when reward is expected (Syed et al., 2016). It is highly likely that the details depend on species, on the specifics of the task and reward involved, on brain region and on exactly what is being measured. Some dopamine signals may be localized, with others more distributed (Berke, 2018). Even the classification of dopamine signals into distinct tonic and phasic components, while supported by prior experimental evidence (Floresco, West, Ash, Moore, & Grace, 2003; Goto, Otani, & Grace, 2007; Niv et al., 2007), has been recently called into question (Berke, 2018; Hamid et al., 2016; Schultz, 2016). Thus, more nuanced modeling of dopamine dynamics, possibly including modulation of dopamine release by striatal cholinergic interneurons (Zhou, Liang, & Dani, 2001), will be an important direction for future work on corticostriatal plasticity and credit assignment.

Beyond adding more biological detail to future models, another natural step will be the consideration of additional reward scenarios. Modeling and experiments on reinforcement learning and action selection often involve probabilistic paradigms, in which different reward probabilities are linked with different actions (Dunovan et al., 2019; Frank et al., 2015; Vich et al., 2020), and it will be essential to confirm that the effectiveness of any solution to the credit assignment problem extends to such settings. The similarity of reward probabilities across options gives a measure of conflict between these choices, while the frequency of changes in contingencies gives a measure of volatility inherent in the situation. It would be interesting to study how the CBGT network with corticostriatal plasticity can encode uncertainty and implement belief updating (Nassar, Wilson, Heasly, & Gold, 2010) as needed to handle conflict and volatility. Another form of conflict could be internal. Some studies have suggested that exploratory actions can occur when cortical signals override the basal ganglia selection mechanisms (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Wilson, Geana, White, Ludvig, & Cohen, 2014; Zajkowski, Kossut, & Wilson, 2017). In theory, sustained cortical activation resulting when a rewarding outcome follows such a directed exploratory action could make that action more likely in the future. Hence, the framework that we suggest could be used to study how exploratory actions can become typical responses. Models linking spiking CBGT dynamics to features of the decision-making process and reinforcement learning will also provide a useful tool for generating predictions about changes in CBGT features underlying altered behavior in pathological states accompanied by changes in action-selection tendencies (Frank, Scheres, & Sherman, 2007; Moustafa, Cohen, Sherman, & Frank, 2008; Wei, Rubin, & Wang, 2015).

Put all together, our review of the literature shows strong progress in understanding how CBGT pathways contribute to reinforcement learning. Historically, the observation that actions may be followed by striatal dopamine signals that scale with reward prediction error and that influence corticostriatal plasticity offered an elegant mechanism for learning to select rewarding actions. This framework proved to be appealing to computational modelers and has featured in theoretical work focusing on issues such as the exploitation–exploration trade-off, the changes in decision-making arising in basal ganglia pathologies and the roles of specific basal ganglia components in the decision-making process. These studies led to the realization that effective learning requires a representation of synaptic eligibility to ensure that credit, in the form of synaptic plasticity, would be localized to those synapses actually involved in making selections that lead to rewards. We have shown how the field is collectively converging on a normative model of the synaptic-level plasticity mechanisms that implement reinforcement learning. Our simulations demonstrate that the experimental observation of sustained cortical activity corresponding specifically to selected actions offers a parsimonious solution to the credit assignment problem that had been unresolved in previous modeling efforts. Moving forward, including this effect may allow future models of CBGT contributions to learning and decision-making to yield better predictions of behavioral or physiological effects and to refine our understanding of the fundamental underlying computations performed by CBGT pathways.

ACKNOWLEDGEMENTS

The authors thank Kyle Dunovan for initial feedback on the network design and structure reported in this paper.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

AUTHOR CONTRIBUTIONS

JR contributed to conception, literature review, model development, interpretation of results and writing. CV contributed to conception, literature review, model development, coding, simulation and writing. MC and KN performed coding and simulation. TV contributed to conception, literature review, model development, interpretation of results and writing.

ORCID

Jonathan E. Rubin https://orcid.org/0000-0002-1513-1551
Timothy Verstynen https://orcid.org/0000-0003-4720-0336

REFERENCES

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*(1), 357–381.

Alves da Silva, J., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, *554*, 244–248.

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: Role of the subthalamic nucleus. *Journal of Neuroscience*, *26*(9), 2424–2433.

Baladron, J., Nambu, A., & Hamker, F. H. (2019). The subthalamic nucleus-external globus pallidus loop biases exploratory decisions towards known alternatives: A neuro-computational study. *European Journal of Neuroscience*, *49*, 754–767.

Balasubramani, P., Pragathi, P., Chakravarthy, V. S., Ravindran, B., & Moustafa, A. A. (2014). An extended reinforcement learning model of basal ganglia to understand the contributions of serotonin and dopamine in risk-based decision making, reward prediction, and punishment learning. *Frontiers in Computational Neuroscience*, *8*, 47.

Bariselli, S., Fobbs, W., Creed, M., & Kravitz, A. (2018). A competitive model for striatal action selection. *Brain Research*, *1713*, 70–79. https://doi.org/10.1016/j.brainres.2018.10.009

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioural Neurosciences*, *4*, 170.

Beiser, D. G., Hua, S. E., & Houk, J. C. (1997). Network models of the basal ganglia. *Current Opinion in Neurobiology*, *7*(2), 185–190.

Berke, J. D. (2018). What does dopamine mean?*Nature Neuroscience*, *21*(6), 787–793.

Berns, G. S., & Sejnowski, T. J. (1996). How the basal ganglia make decisions. In A. R. Damasio, H. Damasio, & Y. Christen (Eds.), *Neurobiology of Decision-making* (pp. 101–113). Berlin, Heidelberg: Springer-Verlag.

Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, *10*(1), 108–121.

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125.

Bogacz, R. (2017). Theory of reinforcement learning and motivation in the basal ganglia. *BioRxiv*, 174524.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

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

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

Buede, D. M. (2013). Decision making and decision analysis. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of Operations Research and Management Science*. Boston, MA: Springer US.

Calabresi, P., Picconi, B., Tozzi, A., & Di Filippo, M. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neurosciences*, *30*(5), 211–219.

Chakravarthy, V. S., Joseph, D., & Bapi, R. S. (2010). What do the basal ganglia do? A modeling perspective. *Biological Cybernetics*, *103*(3), 237–253.

Christopoulos, V., Bonaiuto, J., & Andersen, R. A. (2015). A biologically plausible computational theory for value integration and action selection in decisions with competing alternatives. *PLoS Computational Biology*, *11*(3), e1004104.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1585–1599.

Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801–814. https://doi.org/10.1016/j.neuron.2005.01.027

Coallier, É., Michelet, T., & Kalaska, J. F. (2015). Dorsal premotor cortex: Neural correlates of reach target decisions based on a color-location matching rule and conflicting sensory evidence. *Journal of Neurophysiology*, *113*(10), 3543–3573.

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, *482*(7383), 85–88.

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*(3), 337–366.

Cui, G., Jun, S. B., Jin, X., Pham, M. D., Vogel, S. S., Lovinger, D. M., & Costa, R. M. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature*, *494*(7436), 238–242.

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

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876.

Dekleva, B. M., Kording, K. P., & Miller, L. E. (2018). Single reach plans in dorsal premotor cortex during a two-target task. *Nature Communications*, *9*(1), 3556.

DeLong, M. R., & Wichmann, T. (2007). Circuits and circuit disorders of the basal ganglia. *Archives of Neurology*, *64*(1), 20–24.

Ding, L., & Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. *Neuron*, *79*(4), 640–649.

Draglia, V., Tartakovsky, A. G., & Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests. I. asymptotic optimality. *IEEE Transactions on Information Theory*, *45*(7), 2448–2461.

Dreyer, J. K., Herrik, K. F., Berg, R. W., & Hounsgaard, J. D. (2010). Influence of phasic and tonic dopamine release on receptor activation. *Journal of Neuroscience*, *30*(42), 14273–14283.

Dunovan, K., Lynch, B., Molesworth, T., & Verstynen, T. (2015). Competing basal ganglia pathways determine the difference between stopping and deciding not to go. *Elife*, *4*, e08723.

Dunovan, K., & Verstynen, T. (2016). Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in Neuroscience*, *10*, 106. https://doi.org/10.3389/fnins.2016.00106

Dunovan, K., Vich, C., Clapp, M., Verstynen, T., & Rubin, J. (2019). Reward-driven changes in striatal pathway competition shape evidence evaluation in decision-making. *PLoS Computational Biology*, *15*(5), e1006998.

Fife, K. H., Gutierrez-Reed, N. A., Zell, V., Bailly, J., Lewis, C. M., Aron, A. R., & Hnasko, T. S. (2017). Causal role for the subthalamic nucleus in interrupting behavior. *Elife*, *6*.e27689. https://doi.org/10.7554/eLife.27689

Fino, E., Glowinski, J., & Venance, L. (2005). Bidirectional activity-dependent plasticity at corticostriatal synapses. *Journal of Neuroscience*, *25*(49), 11279–11287.

Fino, E., & Venance, L. (2010). Spike-timing dependent plasticity in the striatum. *Frontiers in Synaptic Neuroscience*, *2*, 6.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

Fisher, S. D., Robertson, P. B., Black, M. J., Redgrave, P., Sagar, M. A., Abraham, W. C., & Reynolds, J. N. J. (2017). Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nature Communications*, *8*(1), 334.

Floresco, S. B., West, A. R., Ash, B., Moore, H. G., & Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nature Neuroscience*, *6*, 968–973.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*, 51–72.

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

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, *35*(2), 485–494.

Frank, M. J., Scheres, A., & Sherman, S. J. (2007). Understanding decision-making deficits in neurological conditions: Insights from models of natural action selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1641–1654.

Gallivan, J. P., Logan, L., Wolpert, D. M., & Flanagan, J. R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature Neuroscience*, *19*(2), 320.

Gallivan, J. P., Stewart, B. M., Baugh, L. A., Wolpert, D. M., &Flanagan, J. R. (2017). Rapid automatic motor encoding of competing reach options. *Cell Reports*, *18*(7), 1619–1626.

Gillies, A., & Arbuthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, *15*(5), 762–770.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(30), 535–561.

Gonon, F. (1997). Prolonged and extrasynaptic excitatory action of dopamine mediated by d1 receptors in the rat striatum in vivo. *Journal of Neuroscience*, *17*(15), 5972–5978.

Goto, Y., Otani, S., & Grace, A. A. (2007). The yin and yang of dopamine release: A new perspective. *Neuropharmacology*, *53*(5), 583–587.

Gurney, K. N., Humphries, M. D., & Redgrave, P. (2015). A new framework for cortico-striatal plasticity: Behavioural theory meets in vitro data at the reinforcement-action interface. *PLoS Biology*, *13*(1), e1002034.

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

Gurney, K., Prescott, T. J., & Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, *84*(6), 411–423.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Weele, C. M. V., … Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, *19*(1), 117–126.

Hernández-López, S., Tkatch, T., Perez-Garci, E., Galarraga, E., Bargas, J., Hamm, H., & Surmeier, D. J. (2000). D2 dopamine receptors in striatal medium spiny neurons reduce l-type Ca^{2+} currents and excitability via a novel PLCB1–IP3–calcineurin-signaling cascade. *Journal of Neuroscience*, *20*(24), 8987–8995.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309.

Hong, S., & Hikosaka, O. (2011). Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. *Frontiers in Behavioral Neuroscience*, *5*, 15.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, 9.

Humphries, M. D., Lepora, N., Wood, R., & Gurney, K. (2009). Capturing dopaminergic modulation and bimodal membrane behaviour of striatal medium spiny neurons in accurate, reduced models. *Frontiers in Computational Neuroscience*, *3*, 26.

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

Izhikevich, E. M. (2002). Resonance and selective communication via bursts in neurons having subthreshold oscillations. *Biosystems*, *67*(1), 95–102.

Izhikevich, E. M. (2007a). *Dynamical systems in neuroscience*. Cambridge, MA: MIT Press. MR2263523 (2008b:37156).

Izhikevich, E. M. (2007b). Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral Cortex*, *17*(10), 2443–2452.

Kalva, S. K., Rengaswamy, M., Chakravarthy, V. S., & Gupte, N. (2012). On the neural substrates for exploratory dynamics in basal ganglia: A model. *Neural Networks*, *32*, 65–73. Selected Papers from IJCNN 2011.

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288.

Keeler, J., Pretsell, D., & Robbins, T. (2014). Functional implications of dopamine d1 vs. d2 receptors: A 'prepare and select'model of the striatal direct vs. indirect pathways. *Neuroscience*, *282*, 156–175. https://doi.org/10.1016/j.neuroscience.2014.07.021

Klaes, C., Westendorff, S., Chakrabarti, S., & Gail, A. (2011). Choosing goals, not rules: Deciding among rule-based action plans. *Neuron*, *70*(3), 536–548.

Klaus, A., Martins, G. J., Paixao, V. B., Zhou, P., Paninski, L., & Costa, R. M. (2017). The spatiotemporal organization of the striatum encodes action space. *Neuron*, *95*(5), 1171–1180.e7.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature*, *412*(6846), 546.

Kropotov, J. D., & Etlinger, S. C. (1999). Selection of actions in the basal ganglia–thalamocortical circuits: Review and model. *International Journal of Psychophysiology*, *31*(3), 197–217.

Latimer, K. W., Yates, J. L., Meister, M. L. R., Huk, A. C., & Pillow, J. W. (2015). Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, *349*(6244), 184–187.

Lauwereyns, J., Watanabe, K., Coe, B., & Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. *Nature*, *418*(6896), 413–417.

Lee, A. M., Tai, L.-H., Zador, A., & Wilbrecht, L. (2015). Between the primate and 'reptilian' brain: Rodent models demonstrate the role of corticostriatal circuits in decision making. *Neuroscience*, *296*, 66–74.

Li, N., Daie, K., Svoboda, K., & Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, *532*(7600), 459.

Lo, C.-C., & Wang, X.-J. (2006). Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, *9*(7), 956–963.

Mandali, A., Rengaswamy, M., Chakravarthy, V. S., & Moustafa, A. A. (2015). A spiking basal ganglia model of synchrony, exploration and decision making. *Frontiers in Neuroscience*, *9*, 191.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*(1), 22–24.

Mikhael, J. G., & Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLOS Computational Biology*, *12*(9), e1005062.

Miller, R. (1988). Cortico-striatal and cortico-limbic circuits: A two-tiered model of learning and memory functions. In H. J. Markowitsch (Ed.), *Information processing by the brain: Views and hypotheses from a cognitive-physiological perspective* (pp. 179–198). Bern, Switzerland: Huber.

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425.

Moustafa, A. A., Cohen, M. X., Sherman, S. J., & Frank, M. J. (2008). A role for dopamine in temporal decision making and reward maximization in parkinsonism. *Journal of Neuroscience*, *28*(47), 12294–12304.

Nambu, A. (2004). A new dynamic model of the cortico-basal ganglia loop. *Progress in Brain Research*, *143*, 461–466.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, *191*(3), 507–520.

Nomoto, K., Schultz, W., Watanabe, T., & Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *Journal of Neuroscience*, *30*(32), 10692–10702.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*(3), 267–273.

Parent, A., & Hazrati, L.-N. (1995). Functional anatomy of the basal ganglia. i. the cortico-basal ganglia-thalamo-cortical loop. *Brain Research Reviews*, *20*(1), 91–127.

Parker, J. G., Marshall, J. D., Ahanonu, B., Wu, Y.-W., Kim, T. H., Grewe, B. F., … Schnitzer, M. J. (2018). Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. *Nature*, *557*(7704), 177.

Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *Journal of Neuroscience*, *31*(19), 7083–7088.

Pawlak, V., & Kerr, J. N. (2008). Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *Journal of Neuroscience*, *28*(10), 2435–2446.

Peak, J., Hart, G., & Balleine, B. W. (2019). From learning to action: The integration of dorsal striatal input and output pathways in instrumental conditioning. *European Journal of Neuroscience*, *49*(5), 658–671.

Pennartz, C., Ameerun, R., Groenewegen, H., & Lopes da Silva, F. (1993). Synaptic plasticity in an in vitro slice preparation of the rat nucleus accumbens. *European Journal of Neuroscience*, *5*(2), 107–117.

Perrin, E., & Venance, L. (2019). Bridging the gap between striatal plasticity and learning. *Current Opinion in Neurobiology*, *54*, 104–112.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-Based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation*, *24*, 1186–1229.

Richfield, E. K., Penney, J. B., & Young, A. B. (1989). Anatomical and affinity state comparisons between dopamine d1 and d2 receptors in the rat central nervous system. *Neuroscience*, *30*(3), 767–777.

Rusu, S. I., & Pennartz, C. M. (2020). Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized cortico-basal ganglia systems. *Hippocampus*, *30*(1), 73–98.

Sauerbrie, B., Guo, J.-Z., Cohen, J., Mischiati, M., Guo, W., Kabra, M., … Hantman, A. (2020). Cortical pattern generation during dexterous movement is input-driven. *Nature*, *577*(7790), 386–391.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Reviews Neuroscience*, *17*, 183–195.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*(12), 4595–4610.

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

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, *63*(3), 607–624. https://doi.org/10.1152/jn.1990.63.3.607

Sen-Bhattacharya, B., James, S., Rhodes, O., Sugiarto, I., Rowley, A., Stokes, A. B., … Furber, S. B. (2018). Building a spiking neural network model of the basal ganglia on spinnaker. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 823–836.

Shan, Q., Ge, M., Christie, M. J., & Balleine, B. W. (2014). The acquisition of goal-directed actions generates opposing plasticity in direct and indirect pathways in dorsomedial striatum. *Journal of Neuroscience*, *34*(28), 9196–9201.

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848–851.

Shindou, T., Shindou, M., Watanabe, S., & Wickens, J. (2019). A silent eligibility trace enables dopamine-dependent synaptic plasticity for reinforcement learning in the mouse striatum. *European Journal of Neuroscience*, *49*(5), 726–736.

Smeets, W. J., Marin, O., & Gonzalez, A. (2000). Evolution of the basal ganglia: New perspectives through a comparative approach. *The Journal of Anatomy*, *196*(4), 501–517.

Surmeier, D. J., Shen, W., Day, M., Gertler, T., Chan, S., Tian, X., & Plotkin, J. L. (2010). The role of dopamine in modulating the structure and function of striatal circuits. *Progress in Brain Research*, *183*, 148–167.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (1st ed.). Cambridge, MA: MIT Press.

Syed, E. C. J., Grima, L. L., Magill, P. J., Bogacz, R., Brown, P., & Walton, M. E. (2016). Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nature Neuroscience*, *19*, 34–36.

Tecuapetla, F., Jin, X., Lima, S. Q., & Costa, R. M. (2016). Complementary contributions of striatal projection pathways to action initiation and execution. *Cell*, *166*(3), 703–715.

Tecuapetla, F., Matias, S., Dugue, G. P., Mainen, Z. F., & Costa, R. M. (2014). Balanced activity in basal ganglia projection pathways is critical for contraversive movements. *Nature Communications*, *5*, 4315.

Thomas, M. J., Malenka, R. C., & Bonci, A. (2000). Modulation of long-term depression by dopamine in the mesolimbic system. *Journal of Neuroscience*, *20*(15), 5581–5586.

Thurley, K., Senn, W., & Luscher, H.-R. (2008). Dopamine increases the gain of the input-output response of rat prefrontal pyramidal neurons. *Journal of Neurophysiology*, *99*(6), 2985–2997.

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

Vich, C., Dunovan, K., Verstynen, T., & Rubin, J. (2020). Corticostriatal synaptic weight evolution in a two-alternative forced choice task: A computational study. *Communications in Nonlinear Science and Numerical Simulation*, *82*, 105048.

Wei, W., Rubin, J. E., & Wang, X.-J. (2015). Role of the indirect pathway of the basal ganglia in perceptual decision making. *Journal of Neuroscience*, *35*(9), 4052–4064.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, *345*(6204), 1616–1620

Yartsev, M. M., Hanks, T. D., Yoon, A. M., & Brody, C. D. (2018). Causal contribution and dynamical encoding in the striatum during evidence accumulation. *Elife*, *7*, e34929.

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife*, *6*, e27430.

Zhang, Y., Fisher, S., Oswald, M., Wickens, J., & Reynolds, J. (2019). Coincidence of cholinergic pauses, dopaminergic activation and depolarization drives synaptic plasticity in the striatum. *bioRxiv*, 803536.

Zhou, F.-M.,Liang, Y., & Dani, J. A. (2001). Endogenous nicotinic cholinergic activity regulates dopamine release in the striatum. *Nature Neuroscience*, *4*, 1224–1229.

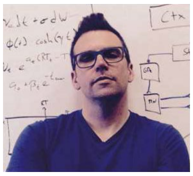
AUTHOR BIOGRAPHIES

Jonathan E. Rubin, Ph.D. Jonathan Rubin was raised by feral robots, where he first acquired his passion for math. He received his bachelors in Psychohistory from the College of Terminus in 1991, his doctorate in Artificial Unintelligence from Skynet University in 1996, and has been a professional cat herder in the Department of Mathematics at the University of Pittsburgh since 2000.

Catalina Vich, Ph.D. Catalina Vich was first officer on the unheralded nautical expedition that discovered the lost city of Atlantis deep in the waters off the coast of Kyrgyzstan. Distracted by a compelling episode of "Alf" during their return trip, Cati and her shipmates were blown severely off course and were shipwrecked on the island of Mallorca. Cati tried to escape by posing as an inebriated tourist and boarding a departing cruise ship but was recognized as a phony due to her lack of extreme sunburn. Resigned to her fate, Cati completed a PhD in Mathematics and has been studying mathematical neuroscience ever since.

Matthew Clapp, B.S. In 2019, Matthew Clapp uploaded a service patch to his consciousness that simultaneously upgraded his Dungeons & Dragons playing skills and awarded him a bachelor in Biomedical Engineering from the University of South Carolina. Unbeknownst to the developers of the patch, Matthew had installed a separate partition in 2017 that allowed him to do research in computational neuroscience. Matthew is currently defragging his memory in order to expand this partition and install a dual boot of the Ph.DOS 1.0 operating system in computational neuroscience from Carnegie Mellon University.

Kendra Noneman After exposure to vibranium radiation, Kendra Nonemean developed superpowers that gave her both the strength and agility of a professional athlete and the intelligence of a certified genius. Kendra is currently finishing mastering material sciences at Boise State University and will move on to start work on the first of her eventual seven Ph.D's in the Fall of 2020 (while also training to arm wrestle Thanos for his golden gauntlet).



Timothy Verstynen, Ph.D. In Tim Verstynen's earliest memories, he is a teenager emerging from a hollow under the bushes on the edge of the Carnegie Mellon campus. Overwhelmed by the bustle of students in day-glo neon T-shirts, he stands transfixed, unable to decide on his next move, whereupon he is knocked down by a passing rollerblader and barely scrambles to safety in the Psychology Department. Tim spent the next two decades lurking in the backs of classrooms, surviving largely on discarded pizza crusts and old coffee, until the faculty realized that he had absorbed the entire curriculum and hired him on the spot. The trauma of his fateful moment of indecision led to Tim's obsession with action selection, which continues today.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Rubin JE, Vich C, Clapp M, Noneman K, Verstynen T. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *Eur J Neurosci*. 2020;00:1–20. https://doi.org/10.1111/ejn.14745