A biologically motivated synthesis of accumulator and reinforcement-learning models for describing adaptive decision-making

by

Kyle Dunovan

B.S., University of Nebraska at Omaha, 2011

Submitted to the Graduate Faculty of

The Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

THE DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kyle Dunovan

It was defended on

December 6, 2016

and approved by

Beatriz Luna, Professor, Psychiatry

Jonathan Rubin, Professor, Mathematics

Natasha Tokowicz, Associate Professor, Psychology

Committee Chair: Julie Fiez, Professor, Psychology

Committee Co-Chair: Timothy Verstynen, Assistant Professor, Psychology

A biologically motivated synthesis of accumulator and reinforcement-learning models for describing adaptive decision-making

Kyle Dunovan, PhD

University of Pittsburgh, 2016

Cognitive process models, such as reinforcement learning (RL) and accumulator models of decision-making, have proven to be highly insightful tools for studying adaptive behaviors as well as their underlying neural substrates. Currently, however, two major barriers exist preventing these models from being applied in more complex settings: 1) the assumptions of most accumulator models break down for decisions involving more than two alternatives; 2) RL and accumulator models currently exist as separate frameworks, with no clear mapping between trial-to-trial learning and the dynamics of the decision process. Recently I showed how a modified accumulator model, premised off of the architecture of cortico-basal ganglia pathways, both predicts human decisions under uncertainty and evoked activity in cortical and subcortical control circuits. Here I present a synthesis of RL and accumulator models that is motivated by recent evidence that the basal ganglia acts as a site for integrating trial-wise feedback from midbrain dopaminergic neurons with accumulating evidence from sensory and associative cortices. I show how this hybrid model can explain both adaptive go/no-go decisions and multi-alternative decisions in a computationally efficient manner. More importantly, by parameterizing the model to conform to various underlying assumptions about the architecture and physiology of basal ganglia pathways, model predictions can be rigorously tested against observed patterns in behavior as well as neural recordings. The result is a biologically-constrained and behaviorally

tractable description of trial-to-trial learning effects on decision-making among multiple alternatives.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First and foremost, to my advisor, Dr. Timothy Verstynen. You've been an incredible source of motivation, knowledge, and personal support over the course of my graduate training and I consider myself supremely lucky to have you as a mentor, colleague, and friend. So, sincerely. Thank you. Gina, Patrick, and Kevin. You will all be lifelong friends. I can't thank you enough for everything that you've done for me. Finally, to my incredible family, you are awesome and I would never have been able to make it to where I am without your support. Dad, this dissertation is for you.

1.0 INTRODUCTION

The flexibility of behavioral control is a testament to the brain's capacity for dynamically resolving uncertainty in the interest of goal-directed action. At the intersection of biology and psychology, the field of cognitive neuroscience is tasked with the challenge of describing the link between complex, adaptive behaviors and the neural processes from which they arise. Recognizing the need for different levels of analysis, David Marr (1982) famously proposed a three-tier system for characterizing neurocognitive phenomena including higher level, computational models of behavior, algorithmic models of the underlying cognitive mechanisms, and lower-level models of neural implementation (Figure 1). Since Marr first proposed the need for different levels of analysis, significant progress has been made in understanding the computational, cognitive, and neural mechanisms underlying both basic decision-making (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Mulder, van Maanen, & Forstmann, 2014; Park, Meister, Huk, & Pillow, 2014; Ratcliff & McKoon, 2008) and reinforcement learning (Cockburn, Collins, & Frank, 2014; Frank, Seeberger, & O'reilly, 2004; Lee, Seo, & Jung, 2012; Schultz, Dayan, & Montague, 1997; Schultz & Dickinson, 2000).

At the highest and most abstract of Marr's three levels, Computational theories seek to identify and characterize the fundamental problem being solved – what are the goals of the system? At the middle, Cognitive models seek to explain the psychological process used to achieve these computational goals – what is the algorithm that converts some input into

behaviorally meaningful output? Finally, the lowest and most concrete level seeks to describe how cognition is physically realized by the brain – how are cognitive algorithms implemented in a physical substrate? The studies conducted in this dissertation are expressly intended to address questions at the cognitive level, but are motivated by what is known about the underlying neural systems responsible for adaptive decision-making. Cognitive models of decision-making predominantly fall within the broader class of accumulation-to-bound models Figure 2A, in which a decision is computed by accumulating the evidence until a threshold is met and a choice can be made (Brown & Heathcote, 2008; Ratcliff & Smith, 2004; Ratcliff, Smith, Brown, & McKoon, 2016; Trueblood, Brown, & Heathcote, 2014; Wagenmakers, van der Maas, & Grasman, 2007). Accumulator models are somewhat unique in that they have been used to simultaneously describe behavior as well as the dynamics of choice-related neural activity (Churchland, Kiani, & Shadlen, 2008; Marshall et al., 2009; Murakami, Vicente, Costa, & Mainen, 2014; Polanía, Krajbich, Grueschow, & Ruff, 2014; Shadlen & Shohamy, 2016; M. N. Shadlen & Newsome, 1996). There are, however, several limitations of accumulator models that vastly limit the scale at which they can be used to make inferences about decision-making in the real world. For instance, most accumulator models are restricted to describing decisions between two alternatives and assume that parameters are fixed across time. Thus, these models provide a useful resource for describing the basic building blocks of the decision process, but fall short of describing how the decision process changes as a function of experience or in response to feedback from the environment.

Critically, these particular limitations are addressed by a different cognitive framework, called reinforcement learning (RL; Figure 2B). Basic RL theory prescribes an algorithm for learning to distinguish between actions that are "good" (e.g., actions that propel you closer to

some goal state) from those that are "bad" (e.g., actions and propel you in the opposite direction or, at best, fail to return the expected results). The major limitations of RL are complementary to those of accumulator models. RL offers insight into why we consider some actions better than others without offering any detailed insights into *how* we actually choose to execute good actions and avoid bad ones; whereas the tradeoff is flipped for accumulator models.

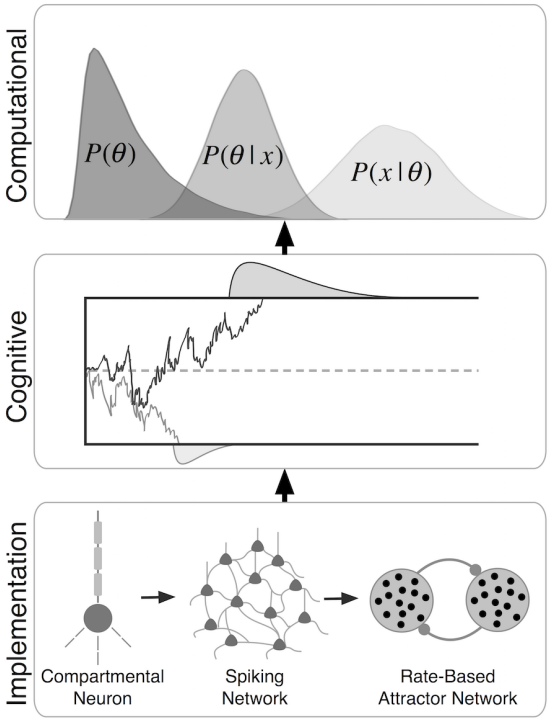

Figure 1. Marr's levels of analysis.

Models at the computational level (top) are intended to capture the computational goals of a system, visualized here as a statistical model of Bayesian inference which posits that internal prior beliefs P(θ) are updated by computing the posterior probability P(θ|x) of over alternative hypotheses $\theta \in \{\theta_i, \ldots \theta_n\}$ given their respective likelihoods P(x|θ). The cognitive level (middle), also referred to as the representational or algorithmic level, aims to describe the psychological process responsible for generating behavioral phenomena. The middle schematic depicts the drift-diffusion model (DDM; Ratcliff (1978); see Ratcliff et al. (2016) for a recent review), a popular model of binary choice based on Wald's Sequential Probability Ratio Test, the Bayes optimal solution for choosing between two hypotheses in the shortest amount of time given an accepted level of error. Models at the implementation level (bottom) intend to describe how computational principles and cognitive algorithms arise from a physical (neural) substrate. Such models range widely in scope, from compartmentalized models of individual neurons to local circuit interactions between populations of neurons to the functional dynamics across entire networks. It is convenient to discuss and represent these levels as being separated by discrete boundaries, however the modern view has shifted towards thinking of these levels as existing on a gradient that ranges from abstract descriptions to lower level biological models.

Traditionally, neurocognitive theories of decision-making and RL have proceeded in parallel and have mostly focused on non-overlapping areas of the brain. However, recent efforts to integrate cognitive modeling with experimental neuroscience have uncovered promising links between these theoretical models and a subcortical network called the basal ganglia (BG; see Figure 3). Over the past decade it has become increasingly clear that, in addition to its well-known role in guiding feedback-dependent learning, the BG is also a critical node in the larger decision-making network (Balleine, Delgado, & Hikosaka, 2007; Dunovan & Verstynen, 2016; Lo & Wang, 2006), acting as a critical way-station for integrating decision-related inputs from cortex and the midbrain dopaminergic signals that drive RL (Bogacz & Gurney, 2007; Cockburn et al., 2014; Forstmann, Anwander, et al., 2010; Frank et al., 2015; Gurney, Humphries, & Redgrave, 2015; Ratcliff & Frank, 2012). Thus, there exists a clear, well defined biological substrate for regulating adaptive goal-directed behavior; however, how RL and decision-making algorithms at the cognitive level emerge from the pathway-level organization of cortico-BG networks remains to be explicated.

Here, I present a series of theoretical and empirical experiments designed to evaluate the utility of encoding action uncertainty as a dynamic competition between opposing control pathways in this network that facilitate (i.e., a Believer) or suppress (i.e., a Skeptic) a decision. When combined with principles of feedback-dependent learning prescribed by RL (Sutton, Barto, & Book, 1998), I show that this Believer-Skeptic framework offers a biologically motivated and behaviorally tractable account of the neural and cognitive mechanisms underlying adaptive, goal-directed behavior. Based on the outcomes of simulations in section 2.0 , I test the assumptions of this framework against empirical observations in two behavioral experiments.

The first experiment (section 3.0) tests the assumption that control is learned via feedback-dependent adaptation of the drift-rate of evidence accumulation – representing dopaminergic modulation of the balance of direct- and indirect-pathways – showing that this mechanism effectively tunes proactive control to statistical demands of the environment across contexts. In a second experiment (section 4.0), I explore how feedback-dependent changes in the drift-rate accounts for choice behavior in the context of multiple alternatives, using a reaching paradigm inspired by the classic Iowa Gambling Task (IGT), to investigate the interplay between indices of economic strategy and sensorimotor adaptation.

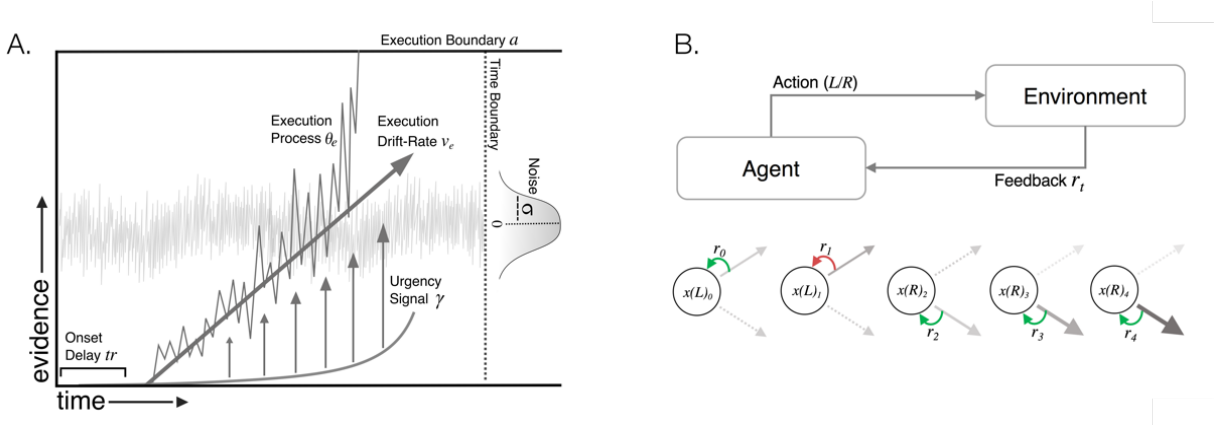

Figure 2. Accumulator and Reinforcement Learning Models.

(A) Parameterization of generic accumulation-to-bound decision model, adapted from Dunovan et al. (2015). **(B)** Schematic of basic Reinforcement Learning framework, showing the effect of positive (green arrows) and negative (red arrows) reinforcement (r_t) on the estimated value $x(.)_t$ of choosing a "left" (L) or "right" (R) action for five sequential trials. Due to negative reinforcement of leftward movement and repeated positive reinforcement of rightward movement, the agent learns to associate moving to the right with a greater value, $x(R)_t > x(L)_t$.

2.0 OVERVIEW OF CORTICO-BASAL GANGLIA CIRCUITRY

2.1 MAJOR PATHWAYS THROUGH THE BASAL-GANGLIA

The BG are thought to be organized as a system of parallel channels, each representing specific task-relevant actions. The canonical model (Albin, Young, & Penney, 1989; Alexander, DeLong, & Strick, 1986) of the BG assumes that the input to each action channel follows two separate pathways to the main output nucleus of the BG, the internal globus pallidus (GPi): a *direct* pathway which facilitates action execution (Figure 3; green) and an *indirect* pathway which suppresses action execution (Figure 3; blue). Cortical projections to distinct subpopulations of medium spiny neurons (MSNs) in the striatum form the inputs to the direct (dMSNs) and indirect (iMSNs) pathways. Activation of the direct pathway suppresses tonic firing in the GPi, thereby facilitating action execution by disinhibiting thalamic activation of primary motor cortex (M1). Conversely activation of the indirect pathway strengthens GPi output via inhibitory connections with the external globus pallidus (GPe) and subthalamic nucleus (STN), thereby suppressing action execution. In contrast to the channel-focused regulation of BG output by the direct and indirect pathways (Albin et al., 1989), activation of the STN through the hyper-direct pathway (Figure 3; red) leads to strong, diffuse activation of the GPi, acting as a general "braking" signal on all action channels (Aron & Poldrack, 2006; Aron, Robbins, & Poldrack, 2014).

Central to the canonical model is the assumption that, for a given action channel, the direct and indirect pathways are parallel and independent until converging in the GPi. In recent years, combined optogenetic and behavioral experiments (Cazorla et al., 2014; Cui et al., 2013; Friedman et al., 2015; Oldenburg & Sabatini, 2015) have strongly challenged this assumption, suggesting that competition between these pathways is fundamental to many BG-mediated behaviors. These experimental techniques have also led to the discovery of novel architectural and physiological properties within the BG that hold promise for resolving disagreements about its role in generating adaptive rather than habitual behavior. For instance, lateral connections have been identified connecting direct and indirect MSNs in the striatum (Taverna, Ilijic, & Surmeier, 2008). Moreover, a significant proportion of direct pathway MSNs send projections to the GPe of the indirect pathway, referred to as bridging collaterals (Cazorla et al., 2014) (Figure 3; green dotted-line). This interaction is compounded by a recently identified feedback loops from GPe back to the striatum (Mallet et al., 2012), delivering widespread inhibition (Silberberg & Bolam, 2015) to both major MSN subtypes and fast-spiking interneurons (FSIs). Along with these more recently discovered structural pathways, the well-known convergence of the direct and indirect pathways at the GPi (Smith et al., 1998b) also implies that these opposing control pathways directly compete for control over BG output.

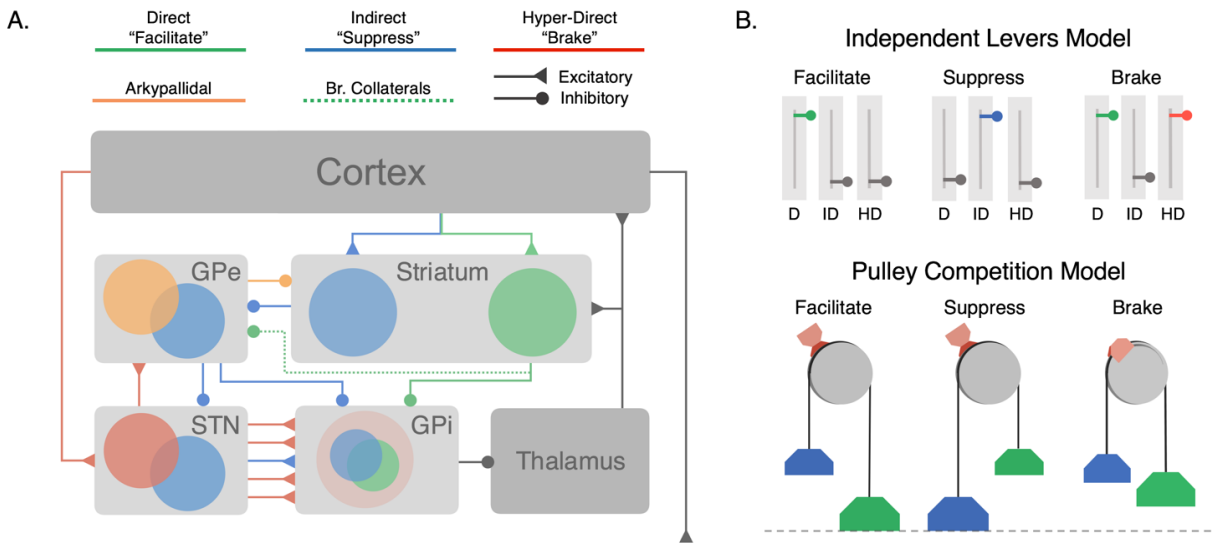

Figure 3. Cortico-basal ganglia pathways and control models

(A) Cortico-BG pathways including three major inputs to the striatal direct (green), indirect (blue) pathways, and the subthalamic hyper-direct (red) pathway. Bridging collaterals (green, dotted) connect the direct pathway to the indirect pathway via projections to the GPe. The arkypallidal pathway (orange) sends inhibitory feedback projections from the GPe to the striatum. Both the direct pathway (cortex-striatum-GPi) and "short" indirect pathway (cortex-striatum-GPe-GPi) form focused projections throughout the network corresponding to individual action channels. The "long" indirect pathway (cortex-striatum-GPe-STN-GPi) and hyper-direct pathway (cortex-STN-GPi) deliver diffuse excitatory inputs to the output nucleus. **(B)** Independent Levers Model (i.e., the canonical model) assumes that the direct (left, green), indirect (middle, blue), and hyper-direct (right, red) pathways are structurally and functionally segregated. Each pathway is operated in isolation for facilitating, suppressing, or braking motor output in the BG. **(C)** Pulley Competition Model (i.e., Believer-Skeptic) assumes that the direct and indirect pathways compete throughout the BG, with the strength of each pathway acting as weights on opposing sides of a pulley. As activation in the direct pathway overpowers that of the indirect pathway, this imbalance accelerates the network toward "facilitation," resulting in an executed action when the difference reaches a critical threshold (dotted line). In the event of a stop cue, the action can be reactively canceled if the pulley brake (red brake pad) is activated before the direct-indirect difference reaches a critical threshold. The accelerating (e.g., nonlinear) dynamics of an imbalanced pulley lead to less efficacious braking when the network is pulled further toward action execution (e.g., longer brake streaks on pulley wheel). This dependency illustrates how proactive modulation of the direct-indirect balance may influence reactive stopping via activation of the hyper-direct pathway. Adapted from Dunovan and Verstynen (2016).

At a conceptual level, the canonical model (Albin et al., 1989) proposes that the cortico-BG pathways act as independent levers (Figure 3B), operating in a mutually exclusive manner for individual actions. This depiction conflicts with the growing body of structural and functional evidence (Cazorla et al., 2014; Cui et al., 2013; Dunovan et al., 2015; Mallet et al., 2016) that these pathways dynamically compete for control over BG output (i.e., whether an action is gated or remains suppressed). For instance, motor output is preceded by co-activation of direct and indirect MSNs in the striatum (Cui et al., 2013). This finding has been interpreted as evidence of a center-surround mechanism (Mink, 1996), where the direct pathway is activated for a target action (e.g., center) and the indirect pathway is activated for competing actions (e.g., surround) (Friend & Kravitz, 2014). This theoretical insight is an important one and a likely candidate mechanism for action-selection in the BG. However, I will argue that this is an incomplete description of indirect pathway contributions to behavior. In contrast with this view, a recent study found that learning to execute goal-directed behavior was associated with opposing plasticity of cortico-striatal synapses, increasing the excitability of direct MSNs while suppressing the excitability of indirect MSNs (Shan, Ge, Christie, & Balleine, 2014). This outcome suggests, rather than behaving as independent levers, the direct and indirect pathways act as weights on opposing sides of a pulley that bias the network toward a more facilitating or suppressing state for a given action (Figure 3B). Over the course of learning, more weight is added to the direct pathway of sensorimotor mappings that yield positive results or facilitate goal acquisition whereas weight is added to the indirect pathway of aversive mappings (Frank, 2005). Indeed, recent theoretical studies have found that effective action selection *requires* simultaneous activation within the direct and indirect pathways of all action channels, whereas independent

pathway activation leads to simultaneous activation of competing actions or a failure to execute any action at all (Gurney et al., 2015).

The pulley model rests on the assumption that, in addition to supplying general surround inhibition, the indirect pathway is also capable of selectively opposing action-specific signals in the direct pathway. Several lines of evidence support this duality.. First, projections from the GPe (short indirect) form proximal synapses on the soma of target cells in the GPi whereas inputs from the STN (long indirect/hyperdirect) form distal dendritic synapses with broder patches of cells (Bolam, Hanley, Booth, & Bevan, 2000; Parent & Hazrati, 1995; Smith, Bevan, Shink, & Bolam, 1998b). These more focal projections to the BG output via the short indirect pathway are mirrored by projections originating from striatal direct pathway cells (Smith, Bevan, Shink, & Bolam, 1998a). Thus, it is plausible that at least one of the functions of the short indirect pathway is to selectively oppose the facilitating influence of the direct pathway on single actions. Recent investigation into the functional topology of striatal MSNs has revealed substantial spatial clustering of functionally correlated MSNs, each cluster comprising simultaneously active direct and indirect pathway neurons (Barbera et al., 2016). Predictive modeling showed that the combined activity of both dMSN and iMSN-contributors to cluster dynamics tracked locomotive behavior in mice more reliably and that these predictions declined in models that failed to account for spatial clustering or when focusing on a single pathway. Finally, dopaminergic feedback to the striatum has opposing effects on recently active cells in the direct and indirect pathways, reinforcing dMSNs while suppressing iMSNs. Given that both pathways are active prior to movement and that the combined activation of these pathways better accounts for active locomotion than activity in the direct pathway alone (see also Yttri &

Dudman (2016)), the phasic bursts and dips in striatal dopamine should lead to action-specific plasticity in both pathways.

The pulley model is also consistent with the putative role of BG circuitry in RL, storing and updating the relative value of alternative actions via feedback-dependent weighting of cortico-striatal synapses (Bogacz & Larsen, 2011; Frank et al., 2004). This feedback-dependent plasticity provides a critical bridge between prior experience and the moment-to-moment accumulation of evidence for choosing between alternative actions.

2.2 BASAL-GANGLIA CONTRIBUTIONS TO DECISION-MAKING

Based on this evidence that action uncertainty modulates the competitive accumulation of two opposing signals, we recently proposed a novel recapitulation of the role of BG pathways in decision making (Dunovan & Verstynen, 2016). Rather than viewing BG pathways as independent, cortically operated control levers for gating motor output, we propose that the BG encodes action uncertainty through a dynamic competition between populations of action facilitating (i.e., Believer) and suppressing (i.e., Skeptic) units. In this way, uncertainty is resolved on a moment-by-moment basis depending on the instantaneous state of direct and indirect pathways. Because the default state of the BG is heavily motor-suppressing (i.e., the null hypothesis of BG decisions is to suppress actions), the burden of proof falls on the Believer to present sufficient evidence for selecting and executing a particular action. This competition can be viewed as one way that for the BG to implement a *dynamic* threshold on gradually accumulating decision evidence. The Believer-Skeptic framework presented here assumes that cortico-BG pathways implement a decision threshold as a dynamic competition of action

facilitating and suppressing network states. While I propose this to be a more neurally plausible mechanism of threshold implementation than that presented in the DDM, this is not to say that model abstraction in the DDM is not useful. In fact, it is necessary for developing quantitative theories that can be meaningfully parameterized at cognitive and behavioral levels of description. In order for these models to be applied to neural data there must be an appreciation for the mapping between cognitive parameters and the more complex neural processes that they represent.

Within the standard DDM, 'competition' is inherently captured by the accumulating decision process, where each step up or down represents the instantaneous evaluation of two competing hypotheses: an action decision and its null alternative. In the context of basic perceptual decisions, stimuli with high signal-to-noise ratio (SNR) produce faster rates of evidence accumulation toward a decision boundary, and are thus recognized faster and more reliably than noisy stimuli. This is an important point to emphasize, as the unidirectional change in the speed and accuracy of decisions is what fundamentally distinguishes a change in drift-rate from a change in the decision threshold in the standard DDM. As hinted at earlier the decision process can instead be reparameterized to reflect different hypotheses regarding the neural processes responsible for integrating contextual information with sensory evidence (Standage, Blohm, & Dorris, 2014). In the Believer-Skeptic framework, contextual information and sensory evidence converge as weighted cortico-striatal inputs to the direct and indirect pathways of a single action channel (Figure 4A). The strong recurrent dynamics within each pathway lead to bistability in the network output (Figure 4B), an important property for implementing a switch between two states (Simen, 2012). Even when the weighted input to each pathway is comparable, small amounts of noise can disrupt the balance enough to cause a state transition

given sufficient self-excitation. As a result, both pathways initially increase their firing rate then diverge as activation in one pathway supersedes and inhibits the other, switching the network toward a 'Go' or 'NoGo' attractor state (Figure 4B). Thus, rather than the sensory driven drift-rate of the DDM, the moment-to-moment competition between alternative hypotheses in the Believer-Skeptic framework is driven by a weighted combination of contextual and sensory information. This form of competition can be seen in Figure 4C, in which Go/NoGo decisions are made by accumulating the output (right panel) of the direct-indirect competition (left panel) under different levels of contextual uncertainty. When action uncertainty is low, the network is accelerated toward a "Go" state (Figure 4B) by stronger activation of the direct pathway, causing a faster accumulation of decision evidence towards a fixed execution threshold. Neurophysiologically, the fixed upper threshold of decision evidence in Figure 4C (right plot) can be conceptualized as the level of pallidal suppression necessary to disinhibit the thalamus so that an action is executed.

We recently proposed a modified accumulator framework motivated by the general control dynamics of the Believer-Skeptic network in Figure 4, where action decisions are executed by accumulating evidence toward a fixed threshold in the presence of dynamic gain (Dunovan et al., 2015). In our so-called dependent process model (DPM; Figure 5B), I found that contextual information (i.e. cued probability of reward) modulates the drift-rate of the execution process. As action uncertainty increases the drift-rate is suppressed, producing a 'no-go' decision when this suppression prevents the decision process from reaching the execution threshold by the trial deadline (Dunovan et al., 2015).

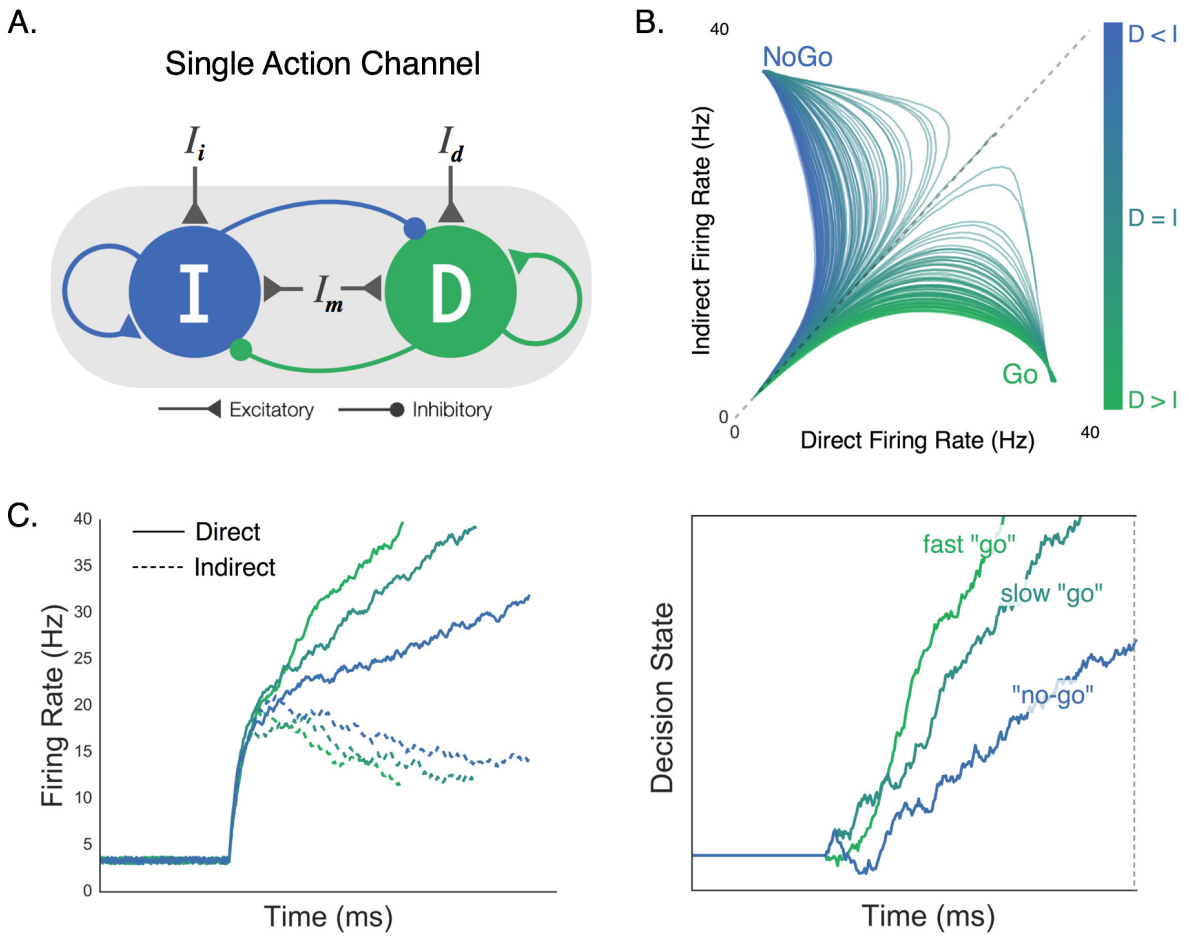

Figure 4. Believer-Skeptic framework as Go-NoGo attractor network

(A) The direct (D) and indirect (I) pathways are modeled as two competing (i.e., mutual inhibition) accumulators with recurrent self-excitation reflecting population attractor dynamics. Selective input to the direct (I_d) and indirect (I_i) pathways is weighted and summed with input from a modulatory (non-selective) population (I_m) which controls the baseline excitability of the network. **(B)** Network state plotted as a function of different ratios of direct and indirect pathway activation. Greater activation of the indirect pathway leads to fast attraction toward a NoGo state (more blue, motor suppressing), whereas greater activation of the direct pathway attracts the network toward a Go state (more green, motor facilitating). **(C)** Left panel: firing rates of direct (solid lines) and indirect pathways (dotted lines) plotted across time for different ratios of input (I_d:I_i). Right panel: accumulation of decision evidence toward an execution threshold, reflecting the normalized difference of the direct and indirect pathways in the left panel. High I_d:I_i ratio accelerates the rate of evidence accumulation, leading to a fast "go" decision (green). As this ratio is reduced (bluish-green), weaker attraction by the direct pathway manifests as a slower rate of accumulation, producing a "no-go" decision when evidence fails to reach threshold by a deadline (blue). Adapted from Dunovan and Verstynen (2016).

Based on the apparent structural overlap of BG pathways in the output nucleus (shown as overlapping red, blue, and green fields in the GPi of Figure 3A), I hypothesized that contextual modulation of competition between direct (i.e., Go) and indirect (i.e., NoGo) pathways should also influence the efficacy of the hyper-direct (i.e., Stop) pathway during reactive action cancellation (Jahfari et al., 2011, 2012; Jahfari, Stinear, Claffey, Verbruggen, & Aron, 2010). Indeed, behavioral fits to RT and choice data in a reactive stop-signal task favored a model in which contextual suppression of the execution drift-rate improves the efficacy of a nested but separate action cancellation process. Collectively, these findings show how the contextual uncertainty associated with a future action is not only critical for making a goal-directed decision about executing that action, but also complements the ability to reactively cancel it based on environmental feedback.

This DPM also captures physiological responses of BG pathways. By integrating the execution process across the trial window, I was able to capture the duration and magnitude of accumulating activity leading up to a decision. Integrating the execution process in this way effectively collapses the decision process into a single measure, similar to how the blood oxygen-level-dependent (BOLD) signal would filter the neural activity generated by attractor network in Figure 4. Consistent with the behavioral fits, I found that contextual modulation of the drift-rate was able to capture the pattern of BOLD activity in the thalamus (the primary output target of the BG pathways) during ‘go’ and ‘no-go’ decisions across varying degrees of uncertainty. This finding is consistent with single-unit recordings of neurons in the macaque motor thalamus which show a similar RT-dependent ramp in firing rate prior to action execution (M. Tanaka, 2007; Masaki Tanaka & Kunimatsu, 2011).

One interpretation of this finding is that pre-action ramping in the thalamus is driven by the differential activation of upstream direct and indirect pathways and thus contextual modulation of this signal occurs by changing the weights of specific cortico-striatal connections or by altering background excitability in the striatum.

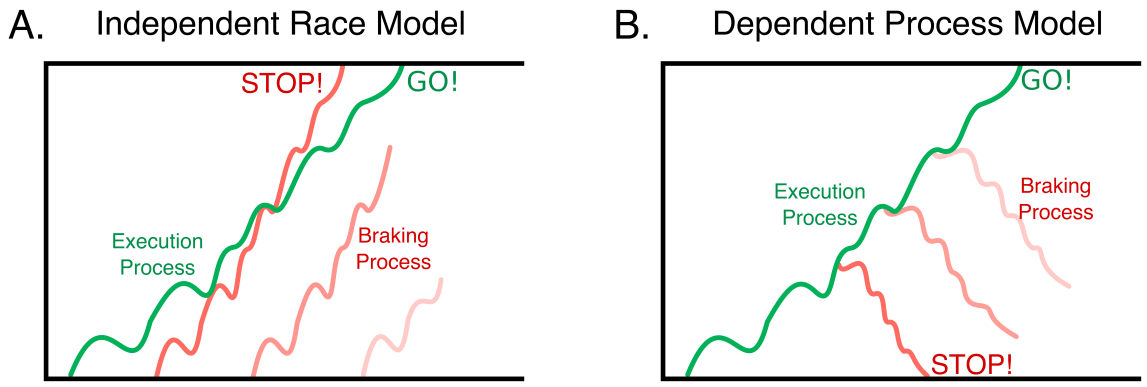

Figure 5. Accumulator Models of Inhibitory Control.

(A) Independent Race Model (IRM), assumes decision to execute an action is represented independently of the decision to "brake" or cancel that action. **(B)** An alternative to the IRM, Dunovan et al. (2015) proposed a Dependent Process Model (DPM) in which the state of the execution decision at the time a stop cue is registered determines initial state of the braking process, making it more difficult to cancel actions closer to the execution boundary.

The hypothesis that the striatum is where contextual information comes to bear on decision evidence is often contrasted with the hypothesis that this is accomplished by the thresholding function of the STN (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010). That is, a change in the slope of thalamic firing rates could be due to decay in the hyper-direct activation of the STN, allowing pallidal suppression by the direct pathway to disinhibit the thalamus at a proportional rate. The distinction between striatal and STN control over decision threshold is a critical one (Bogacz et al., 2010), as these structures have very different input-output motifs that hint at disparate functional roles. The input-output organization of the striatum is thought to be

channel-specific, propagating individual action-commands from cortex to corresponding units in the GPe (indirect) and GPi (direct) segments. The STN, on the other hand, receives converging afferents from cortex and the GPe and delivers more diffuse excitatory drive to the GPi, suggesting this structure modulates the decision threshold in a non-specific manner for all actions under consideration.

In fact, another hypothesis has been proposed for the role of the STN in decision-making that both complements the role of the striatum in the Believer-Skeptic framework and distinguishes the functional relevance of indirect and hyper-direct activation of the STN. Bogacz & Gurney, (2007) presented a neural network model in which the STN normalizes activity in the GPi to accommodate different set sizes of alternative choices. In their model, sensory evidence for each alternative is fed into a corresponding action channel in the striatum in parallel with projections that activate the STN. As a result, the cortico-striatal activation within each individual channel of the GPi (i.e., representing candidate actions 'A', 'B', and 'C', for instance) is represented as a proportion of the evidence for each action relative to the total evidence for all actions under consideration. This model describes the general increase in RT associated with increasing the number of choices to be considered, indicative of a global increase in the threshold for all possible outcomes (Keuken et al., 2015). Another group found that removal of the STN from the network had similar effects on choice RTs as STN deep brain stimulation in treated Parkinson's patients - selectively eliminating the delay in RT for low-probability stimuli (Antoniades et al., 2014).

The proposed thresholding and normalization functions of the STN are complementary with the Believer-Skeptic framework and can be dissociated from the hitherto-proposed role of the direct and indirect competition as a mechanism for encoding action uncertainty. The

normalizing effect of STN output on pallidal inhibition emerges naturally under the assumption that all actions simultaneously engage both the direct and indirect pathways. That is, individual action uncertainty is encoded by the "short" indirect pathway from striatum to GPe and then to channel-specific populations in the GPi (see Figure 3; Schroll & Hamker, 2013) where the indirect pathway converges with action facilitating signals of the direct pathway (Smith et al., 1998b). On the contrary, activation of the "long" indirect pathway, splitting off from GPe to the STN, leads to widespread excitatory increase in GPi firing. Under the assumption that both direct and indirect pathways are active for each action being considered, the net activation through the "long" indirect pathway has a normalizing effect on the basal GPi state, accommodating varied set sizes of alternative actions (Herz, Zavala, Bogacz, & Brown, 2016).

While the long-indirect and hyper-direct pathways likely play an important role in action selection, the within-channel competition of the direct and (short) indirect pathways is ultimately what determines which action is selected. For instance, in the context of a forced-choice perceptual decision, the transition between accumulation and execution is determined by the relative activation of two alternative action channels, each driven by a separate set of competing direct and indirect populations. This process is shown in Figure 6, where an observer must decide whether a noisy field of moving dots contains greater coherent leftward or rightward motion. Critically, a cue is displayed prior to each choice informing the observer which outcome is more likely to be correct on the upcoming trial. Previous work has shown that this predictive information is encoded by a concurrent increase in the baseline activity in the striatum (Forstmann, Brown, Dutilh, Neumann, & Wagenmakers, 2010), contralateral to the expected action, and modulatory regions of cortex, such as orbitofrontal cortex (OFC) and pre-supplementary motor area (preSMA). When the cued probability is valid (i.e., correctly predicts

the subsequent stimulus; Figure 6A) the increase in baseline activity of the corresponding action channel causes the network to become increasingly unstable, leading to faster gating upon descending input from cortical accumulators. However, when the cue is misleading, or invalid (Figure 6B), this destabilization in the cued action channel can lead to an incorrect response despite weak sensory evidence in favor of that choice. This speed-accuracy tradeoff is a widespread phenomenon that pervades many forms of decision-making (Bertucco, Bhanpuri, & Sanger, 2015; Dean, Wu, & Maloney, 2007; Drugowitsch, Deangelis, Angelaki, & Pouget, 2015; Lo, Wang, & Wang, 2015). While numerous studies have found that functional and structural connectivity between preSMA and the striatum predicts individual differences in the speed-accuracy tradeoff (Forstmann, Brown, et al., 2010; Keuken, Langner, Eickhoff, Forstmann, & Neumann, 2014; van Maanen et al., 2011), the underlying mechanism by which modulatory cortical inputs influence action selection in the BG has remained unclear. The example here proposes one such mechanism and highlights an important prediction of the Believer-Skeptic framework: uncertainty associated with individual actions is encoded by the competition between corresponding direct and indirect pathways. Of course, this prediction will need to be more rigorously tested, both experimentally and through the use of more sophisticated computational models of BG circuitry.

The Believer-Skeptic framework provides a novel account for the role of the BG in decision-making, demonstrating the computational utility for encoding action uncertainty in the competition between the direct and indirect pathways. This framework also provides a straightforward interpretation of the different roles of striatal and STN modulation of the decision process. Non-specific background inputs to the striatum can adjust the speed-accuracy tradeoff in favor of faster decision-making by promoting stronger state attraction in response to

descending sensory inputs from cortex. Cortico-striatal mechanisms may also modulate the decision in outcome-specific ways (Majid, Cai, Corey-Bloom, & Aron, 2013) by altering the balance of channel-specific activity in the direct and indirect pathways. This interpretation is consistent with human neuroimaging studies linking cortico-striatal activity to the facilitation of one choice at the expense of choosing another; for instance, by selectively increasing of the drift-rate or baseline evidence for an expected outcome (Dunovan et al., 2015; Forstmann et al., 2010). On the other hand, indirect pathway activation of the STN provides a normalizing constant to BG output by aggregating the activation of multiple action channels into diffuse projections to the GPi (Smith et al., 1998b), whereas hyper-direct activation of the STN modulates the decision indiscriminately, buying time in the interest of accuracy (Forstmann et al., 2012; Frank et al., 2015). In the following, I elaborate on how Believer-Skeptic dynamics of decision-making are complemented by the well-established role of the cortico-striatal circuits in mediating RL.

The Believer-Skeptic framework provides a novel account for the role of the BG in decision-making, demonstrating the computational utility for encoding action uncertainty in the competition between the direct and indirect pathways. This framework also provides a straightforward interpretation of the different roles of striatal and STN modulation of the decision process. Non-specific background inputs to the striatum can adjust the speed-accuracy tradeoff in favor of faster decision-making by promoting stronger state attraction in response to descending sensory inputs from cortex. Cortico-striatal mechanisms may also modulate the decision in outcome-specific ways (Majid et al., 2013) by altering the balance of channel-specific activity in the direct and indirect pathways.

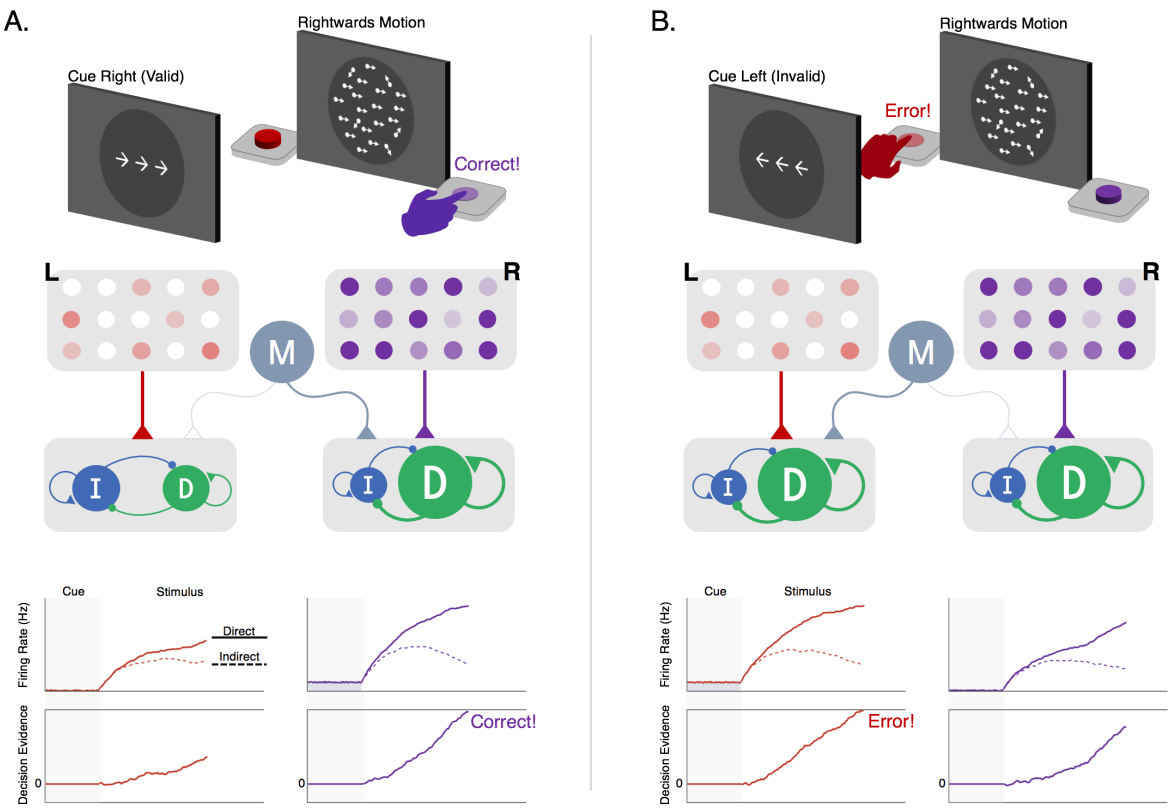

Figure 6. Contextual Modulation of Believer-Skeptic Competition

A) Schematic of cue and stimulus epochs of random dot motion task on trial with valid predictive cue-stimulus combination (Top). Schematic of decision network (Middle): "Left" (L, red) and "Right" (R, purple) motion-selective sensory populations gradually increase activity at a rate proportional to the strength of coherent motion in their preferred direction. Each sensory population sends excitatory input to a corresponding pair of direct and indirect populations representing left- and right-hand actions for reporting leftward and rightward motion decisions, respectively. Sensory inputs activate both pathways but with a bias favoring the direct pathway, reflecting the tendency for sensory inputs to the striatum to form more connections with dMSNs than iMSNs (Wall, De La Parra, Callaway, & Kreitzer, 2013). A modulatory population (M, gray) delivers non-selective excitatory input to the pair of direct and indirect pathways encoding the anticipated action (i.e., action corresponding to the cue-predicted motion direction). Below the network**,** firing rates are plotted for the direct (solid line) and indirect (dotted lines) populations for each choice-hand mapping (bottom-upper). In the bottom-lower panel, plots show the accumulating difference between direct and indirect firing-rates toward an execution threshold. The effect of cued expectations can be seen as an upwards shift in the baseline firing rates of the right-hand direct-indirect network, reflecting anticipatory background inputs from the modulatory population such as preSMA (Forstmann, Anwander, et al., 2010). This increases the excitability of the network, causing a faster separation in the direct-indirect competition and a faster rise-to-threshold in the "Right" decision variable (correct). **(B)** Same task as in the left panel but on a trial with an invalid predictive cue-stimulus combination **(**top**)**. The invalid expectation signal destabilizes the direct-indirect competition, leading to a faster rise-to-threshold of the "Left" decision variable despite receiving less sensory evidence for that option **(**bottom**)**. Adapted from Dunovan and Verstynen (2016).

This interpretation is consistent with human neuroimaging studies linking cortico-striatal activity to the facilitation of one choice at the expense of choosing another; for instance, by selectively increasing of the drift-rate or baseline evidence for an expected outcome (Dunovan et al., 2015; Forstmann et al., 2010). On the other hand, indirect pathway activation of the STN provides a normalizing constant to BG output by aggregating the activation of multiple action channels into diffuse projections to the GPi (Smith et al., 1998b), whereas hyper-direct activation of the STN modulates the decision indiscriminately, buying time in the interest of accuracy (Forstmann et al., 2012; Frank et al., 2015). In the following, I elaborate on how Believer-Skeptic dynamics of decision-making are complemented by the well-established role of the cortico-striatal circuits in mediating RL.

2.3 BASAL-GANGLIA CONTRIBUTIONS TO LEARNING

In the early stages of learning a new skill the brain makes use of past mistakes to improve future performance, incrementally advancing towards a goal-state by trial-and-error. RL models have been highly successful in describing trial-to-trial adaptation in behavior as well as experience-dependent plasticity in putative learning networks in the brain. Basic RL models posit that an agent learns to value more advantages actions by successively comparing the predicted and observed outcome of the last action. This difference, referred to as the prediction error, is scaled by a learning rate which determines the extent to which each observation influences the perceived value of a given stimulus-response mapping. Relatively simple extensions of this algorithm have proven surprisingly adept at accounting for trial-wise choice dynamics across a

wide range of behavioral tasks. More importantly, model-estimated prediction errors have consistently been shown to track with the bursting and pausing of midbrain dopaminergic neurons as well as resulting activity changes in recipient neurons in the striatum, providing a critical link between behavioral and neural signatures of learning (Apicella, Ljungberg, Scarnati, & Schultz, 1991; Schultz & Dickinson, 2000; Surmeier, Ding, Day, Wang, & Shen, 2007).

Electrophysiological studies have consistently found a relationship between the phasic activation of midbrain dopaminergic neurons and the trialwise magnitude of RPEs that mediate RL. For this dopaminergic RPE to be a viable learning signal it must be capable of selectively encouraging rewarded actions and discouraging unrewarded or punished actions. The phasic increase in dopamine following a surprising reward both sensitizes dMSNs and desensitizes iMSNs, making it easier for cortical inputs to quickly execute that action in the future (Hollerman, Tremblay, & Schultz, 1998; Schultz, 2016; Tremblay, Hollerman, & Schultz, 1998) (Hart, Rutledge, Glimcher, & Phillips, 2014; Wiecki & Frank, 2013). By the same token, phasic dips in dopamine following the omission of an expected reward offset the balance in the other direction, requiring stronger or prolonged cortical input to gate the same action in the future (Bahuguna, Aertsen, & Kumar, 2015; Gurney et al., 2015; Marcott, Mamaligas, & Ford, 2014). The bidirectional effect of positive and negative feedback on pathway-specific neural subtypes sheds light on the utility of selecting actions with two opposing pathways instead of a single facilitation pathway (Hart et al., 2014). Indeed, several lines of evidence suggest that dopaminergic modulation of the direct pathway is primarily driven by positive RPEs that facilitate approach-learning, whereas the modulation of the indirect pathway is primarily driven by negative RPEs, facilitating avoidance learning (Cox et al., 2015; Frank, Doll, Oas-terpstra, & Moreno, 2009; Hikida et al., 2013).

In a series of computational experiments, Gurney et al. (2015) recently provided a comprehensive description of the interactions between tonic and phasic fluctuations in striatal dopamine that guide goal-directed action selection. In their neural network model, cortical input from competing sensory populations is sent in parallel to all three cortico-BG pathways representing the sensory-paired actions. Thus, when sensory information is equivocal and cortical input leads to comparable activation in different action channels, the history dependent cortico-striatal weights are what critically determine which of the two actions wins out in the selection process.

The synaptic tuning of these weights by positive and negative RPEs can be naturally incorporated into the Believer-Skeptic decision network shown in Figure 4A – by increasing the sensitivity of the direct and indirect populations following rewarded and punished actions, respectively. Over the course of several trials, the feedback-dependent tuning of synaptic weights leads to faster gating in the network and thus faster rates of evidence accumulation in decision space for higher valued actions. This is captured in Figure 7A where the model gradually learns the relative value of alternative actions based on probabilistic stimulus-reward contingencies from trial-and-error feedback. Similar to the behavioral paradigm used by Frank et al. (2004) the model is presented with a pair of stimuli and must learn to select the stimulus with a higher probability of yielding a reward. Each stimulus is converted into an action by a corresponding pair of direct and indirect nodes that are tuned by corrective feedback signals, simulating the effects of dopaminergic RPE signals on dMSNs and iMSNs. Thus, feedback sensitizes the direct pathway and suppresses the indirect pathway for the optimal choice while shifting the balance in the opposite direction for the alternative, converging on weights that reflect the expected difference in their learned values. In the accumulator model, this manifests as a drift-rate for

each stimulus proportional to its perceived value, leading to a stronger choice bias when deciding between alternatives that are less evenly matched in terms of their expected payout (Figure 7A).

Because in this example the stimulus-action-value associations are probabilistic, a certain amount of exploration is needed in order to optimize the estimated value for each of the two stimuli. In standard RL models, exploratory dynamics are usually facilitated by a single parameter that determines the probability of going with the currently highest-valued option. Here, however, exploration is naturally handled by the stochastic nature of the direct-indirect competition during the decision process. A recent study found that the RT distributions of value-based choices in a perceptual learning experiment were well described by a DDM in which the learned value difference between alternative stimuli determined the drift-rate of accumulation (Frank et al., 2015). This finding adds support to the future hybridization of RL and decision models, suggesting that the behavioral dynamics of value-based choices can be systematically characterized by corrective modulation of a stochastic rise-to-threshold process.

In addition to the phasic dopamine modulations responsible for learning action-value associations, the level of tonic dopamine availability in the striatum has recently been proposed to regulate the tradeoff between exploratory and exploitative learning policies (Humphries, Khamassi, & Gurney, 2012; Kayser, Mitchell, Weinstein, & Frank, 2015). That is, in order to maximize rewards in dynamic environments (with changing response-outcome contingencies), one must balance the time spent exploring the value of novel, potentially high-payoff actions and exploiting historically rewarding actions (Humphries et al., 2012; Keeler, Pretsell, & Robbins, 2014). Put into the context of the Believer-Skeptic framework, explorative states can be thought of as conditions in which the balance is tipped towards the Skeptic such that all action possibilities are uncertain and thus no single decision dominates. In contrast, exploitative states

are those in which the Believer dominates for a single decision, resulting in faster and more precise decisions that preclude alternative actions from being engaged.

Much of the current understanding of the interplay between value-based learning mechanisms and exploitation-exploration tradeoff policies has come from research on song-bird learning (Brainard & Doupe, 2002; Kao, Doupe, & Brainard, 2005). While research on song-bird learning has progressed largely in parallel with the studies of decision-making in the BG, it has been speculated that the two fields are currently moving towards a mutually beneficial junction (Ding & Perkel, 2014). Juvenile song-birds initially learn to sing by mirroring the song of an experienced tutor but over time compose an individualized version of the song by sampling alternate spectral and temporal components of vocalization (Tumer & Brainard, 2007). This is done to improve reproductive success, as females tend to select males with unique songs that can be performed repeatedly with high precision.

Recently, Woolley et al. (2014) found that when practicing in isolation, males express substantially more variability in the spectral and temporal dimensions of song vocalization than when in the presence of a mate. This contextual alternation between exploring alternate song renditions during practice and exploiting a favorite rendition led to systematic differences in the variability of firing in the output of a region called Area X, a homologue of the mammalian BG. The authors proposed that social context led to changes in the tonic level of dopamine available to neurons in the input structure of Area X, similar to the striatum of the BG in mammals, which impacted the amount of exploration or exploitation of the system. Their hypothesis was supported by the observation that striatal connections exhibit a many-to-one convergence onto target cells in the BG output nucleus. Previous work suggests that given this many-to-one motif, enhanced dopaminergic tone would establish a more consistent average level of activation within

a group of striatal units, thus increasing reliability of temporally-locked bursts and pauses of recipient neurons in the output nucleus (Tumer & Brainard, 2007).

Consistent with a dopaminergic regulation between exploitative-explorative policies, several recent computational modeling studies have found that the simulated effects of tonic dopamine level have a marked impact on action variability (Klanker, Feenstra, & Denys, 2013; Morita & Kato, 2014; Yawata, Yamaguchi, Danjo, Hikida, & Nakanishi, 2012). Increasing dopaminergic availability in the striatum leads to a general "Go" bias in the network, due to the inverse effects of dopamine on MSN subpopulations. Furthermore, higher tonic dopamine levels also increases D1 and D2 receptor occupancy so that RPE signals communicated by phasic bursts and pauses in SNc fail to have the same impact on cortico-striatal plasticity (Keeler et al., 2014). Thus, behavior is stabilized to promote exploitation of previously learned associations by facilitating BG throughput that reflects the present weighting scheme at cortico-striatal synapses. In Figure 7B, the population firing rates are shown for different decision policies, all reflecting the same ratio of input to the direct and indirect pathways, but with a change in background levels of tonic dopamine (e.g., background excitation). Increasing dopamine reduces the time constant of evidence accumulation such that learned cortico-striatal weights can be exploited to rapidly accelerate the network toward a "Go" state, with little variability in the RT and outcome of the decision process (Figure 7B). Alternatively, the same levels of cortical input leads to substantially greater trial-to-trial variability in decision behavior when dopamine is scarce, demonstrated by the widening of the RT distribution for decisions made under lower levels of background dopamine.

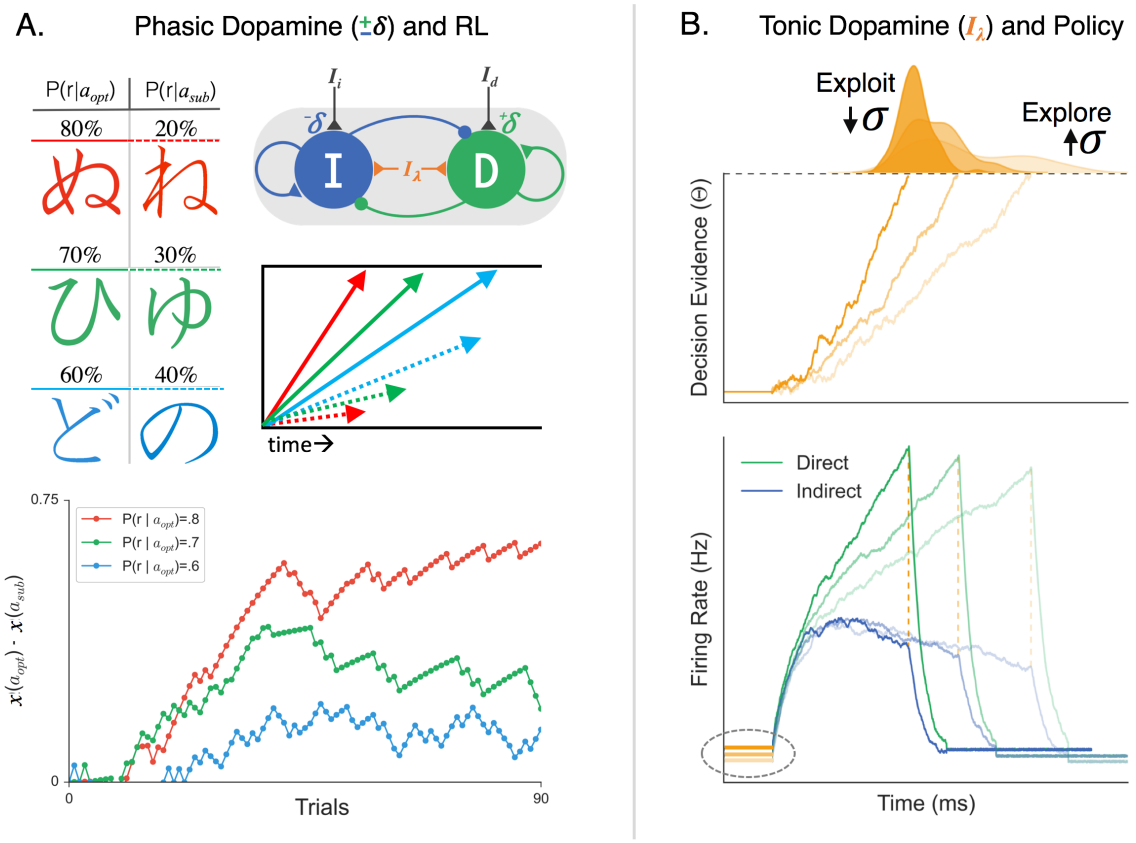

Figure 7. Dopaminergic modulation of value-based decisions

(A) Simulation of probabilistic value-based decision task (upper-left; see Frank et al. (2004)) in which the agent must learn the relative value of two arbitrary stimuli based on trial-and-error feedback. On each trial the agent makes a decision by choosing between a pair of Japanese symbols, one with a higher probability of yielding a reward (left column; chosen with action a_{opt}) than the other (right column; chosen with action a_{sub}). Value-based decisions are simulated as a race-to-threshold between stochastic accumulators, one for each alternative under consideration. Each accumulator reflects the direct-indirect competition within a single action channel (see Figure 4). Both actions start out with equal associated values $x(a_{opt}) = x(a_{sub})$ and thus, equal drift-rates. On each trial the corrective effects of phasic changes in dopamine are simulated by enhancing (depressing) the sensitivity of the direct (indirect) pathway following positive outcomes **(+δ)** and vice-versa following negative outcomes **(−δ)**. In the accumulator model, this learning results in an increase in the drift-rate for a_{opt} (solid arrow) and a decrease in the drift-rate for a_{sub} (dotted arrow), proportional to the difference in their associated value. The bottom panel shows the change in the estimated value difference for alternative actions $x(a_{opt}) - x(a_{sub})$ across 90 simulated trials for the three different probabilistic reward schedules shown in the upper-left. For stimulus pairs with a greater discrepancy in reward probability (i.e., red > green > blue), this leads to earlier separation between drift-rates associated with optimal and suboptimal actions, and thus faster associative value learning. **(B)** Simulated effects of tonic dopamine levels on exploration-exploitation tradeoff. Tonic dopamine levels were simulated by varying the strength of non-specific background inputs (Iλ) in a network with stronger weighting of cortical input to direct than indirect pathway. Bottom panel: the same ratio of cortical input to the direct (green) and indirect (blue) pathways leads to faster gating in the presence higher Iλ (darker colors, increased baseline) compare to when Iλ is low (lighter colors, decreased

baseline). Top panel: Increasing tonic levels of I_λ facilitates exploitation of the current cortico-striatal weights by accelerating evidence accumulation, resulting in faster decisions and reduced trial-to-trial variability in RT. In contrast, behavior is substantially more variable with lower levels of I_λ, promoting an exploration policy. Adapted from Dunovan and Verstynen (2016).

When considered in the context of selecting from multiple actions, the increase in action variability (i.e., wider RT distribution) with reduced levels of tonic dopamine would allow the agent to explore novel, potentially more rewarding, stimulus-action associations. When a sufficiently rewarding association is found or when there is a change in context that demands precision, increasing background dopamine levels would temporarily halt feedback-dependent plasticity to ensure lower variability in performance.

2.4 SUMMARY OF BELIEVER-SKEPTIC FRAMEWORK

An emerging body of evidence points to the BG as a critical site for integrating cortically distributed computations with dopaminergic learning signals in the service of flexible and adaptive behavioral control (Dunovan et al., 2015; Dunovan & Verstynen, 2016; Forstmann et al., 2008; Forstmann, Anwander, et al., 2010; Kayser, Mitchell, Weinstein, & Frank, 2015; Turner & Desmurget, 2010; van Maanen, Fontanesi, Hawkins, & Forstmann, 2016; Wang, Miura, & Uchida, 2013). The Believer-Skeptic framework presented above provides a biologically motivated account of the neural mechanisms underlying the transition from evidence accumulation to action execution. More importantly it lays the foundation for a rich behavioral repertoire when combined with its well-established role of dopamine in reinforcement learning. The simulations presented above touch on a small subset of the behavioral phenomena associated

with BG circuitry and even then, the biological simplicity of the Believer-Skeptic attractor model paints a relatively coarse picture of actual neural implementation. Although this model makes significant compromises in the way of neurobiological detail, the ability to formally simulate behavioral predictions based on abstracted neural dynamics provides a useful tool for exploring the hypothesis space. Obviously the hope is that, by carving out useful connections between brain and behavior, this approach will foster more rapid progress on both ends of the spectrum. In the remaining sections, I test the predictive utility of the Believer-Skeptic framework in two empirical studies that address different components of adaptive behavior.

3.0 ADAPTIVE CONTROL OVER A SINGLE ACTION

3.1 INTRODUCTION

In previous sections (sections 2.2 and 2.3), I introduced a framework for integrating feedback-dependent learning with the stochastic accumulation-to-bound process believed to underlie decision-making. This model is specifically constrained by the known structural and functional properties of cortico-BG circuitry. In contrast with the dominant characterizations of the BG as playing a value-centric or motor-centric role in behavior, the Believer-Skeptic framework proposes that the BG is most critical for integrating and resolving action uncertainty

While previous theories have entertained the notion that the BG is an important structure for guiding behavior in the context of uncertainty, the vast majority have focused on the uncertainty between alternate actions (Bogacz & Gurney, 2007) - i.e., which action is best? This assumption follows from the notion that the BG is composed of *action channels,* each channel containing a direct and indirect pathway from the striatum to the output nucleus reflecting a single action decision. Mink (1996) proposed a center-surround mechanism that would ensure cortical decision outcomes activated the appropriate action without any interference. More recently, however, multiple lines of evidence have suggested that the BG not only influences which action is selected but also how that action is expressed. For instance, model-based neuroimaging studies have found evidence that the striatum acts as a gain dial for adjusting the

urgency in decision formation, thereby speeding action execution. A recent study by (Yttri & Dudman, 2016), in which mice were rewarded for pushing a lever, found that the velocity of the animal's movement could be gradually increased from trial-to-trial by optogenetically stimulating D1-expressing cells in the direct pathway in a closed-loop so that faster movements returned higher levels of stimulation. This velocity-yoked stimulation of neurons in the direct pathway can be seen as an artificial reward signal – producing the same sustained activation in recently fired D1 as the burst of synaptic dopamine triggered by an external reward. Conversely, when velocity-yoked stimulation was directed at D2-expressing cells of the indirect pathway, future movements were gradually slowed, artificially inducing the same penalizing effect on the action as a phasic dip in dopamine. This competitive modulation of movement speed is in line with the feedback-dependent learning and control mechanisms assumed within the Believer-Skeptic framework - specifically, that dopaminergic modulation of the competition between direct and indirect pathways is employed for dialing in goal-relevant dimensions of motor control. Here, I evaluate the predictive utility of this framework in relation to observed patterns of adaptive control behavior in a modified version of the reactive stop-signal task presented in Dunovan et al. (2015).

3.2 METHODS

3.2.1 Participants

Neurologically healthy adult participants (N=75, Mean age 22 years) were recruited from local student population at Carnegie Mellon University. All procedures were approved by the local Institutional Review Board. All subjects were compensated for their participation.

3.2.2 Adaptive Stop-Signal Task

All subjects completed a stop-signal task (N_{trials}=880) in which a vertically moving bar approached a white horizontal target line at the top of the screen (Figure 8A). On 'Go' trials (N_{GO}=600) the subject was instructed to make a key press as soon as the bar crossed the target. The bar always intersected the target line at 520ms after trial onset. On each trial, the bar continued filling upward until a keypress was registered or until reaching the top of the screen, allowing a 680ms window for the subject to make a response. If no response was registered the subject received a penalty of (-100pts). On 'Go' trials where a response was recorded before the 680ms trial response deadline, the subject received a score reflecting the precision of their response time relative to the target intersection, resulting in maximal points when RT=520ms. On 'Stop' trials, the bar would stop and turn red prior intersecting the target line, prompting the subject to withhold their response. Successful and unsuccessful stop trials yielded a reward of +200 points and penalty of -100 points, respectively. On the majority of Stop trials, the stop-signal delay (SSD) - the delay between trial onset and when the bar stopped – was sampled from

a specific probability distribution, as shown in Figure 8B. I refer to these trials as Context trials ($N_{Context}$=200). The type of distribution for Context SSDs was held constant for each group. Context SSD's in the Early and Late groups were sampled from Gaussian distributions with equal variance (σ=35ms), centered at μ_E=250ms and μ_L=350ms, respectively. Context SSDs in the Uniform group were sampled from a uniform distribution spanning a 10-520ms window. In Figure 8B, the sampled SSD times are plotted for a single subject in each of the groups – shown as dashes on a timeline ranging from 0-520ms. Finally, an additional 80 "Probe" Stop trials were included where the bar stopped either at 200, 250, 300, 350, or 400ms (N=16 each) after trial onset. These trials are shown at the bottom of the Figure 8B timeline as red dashes.

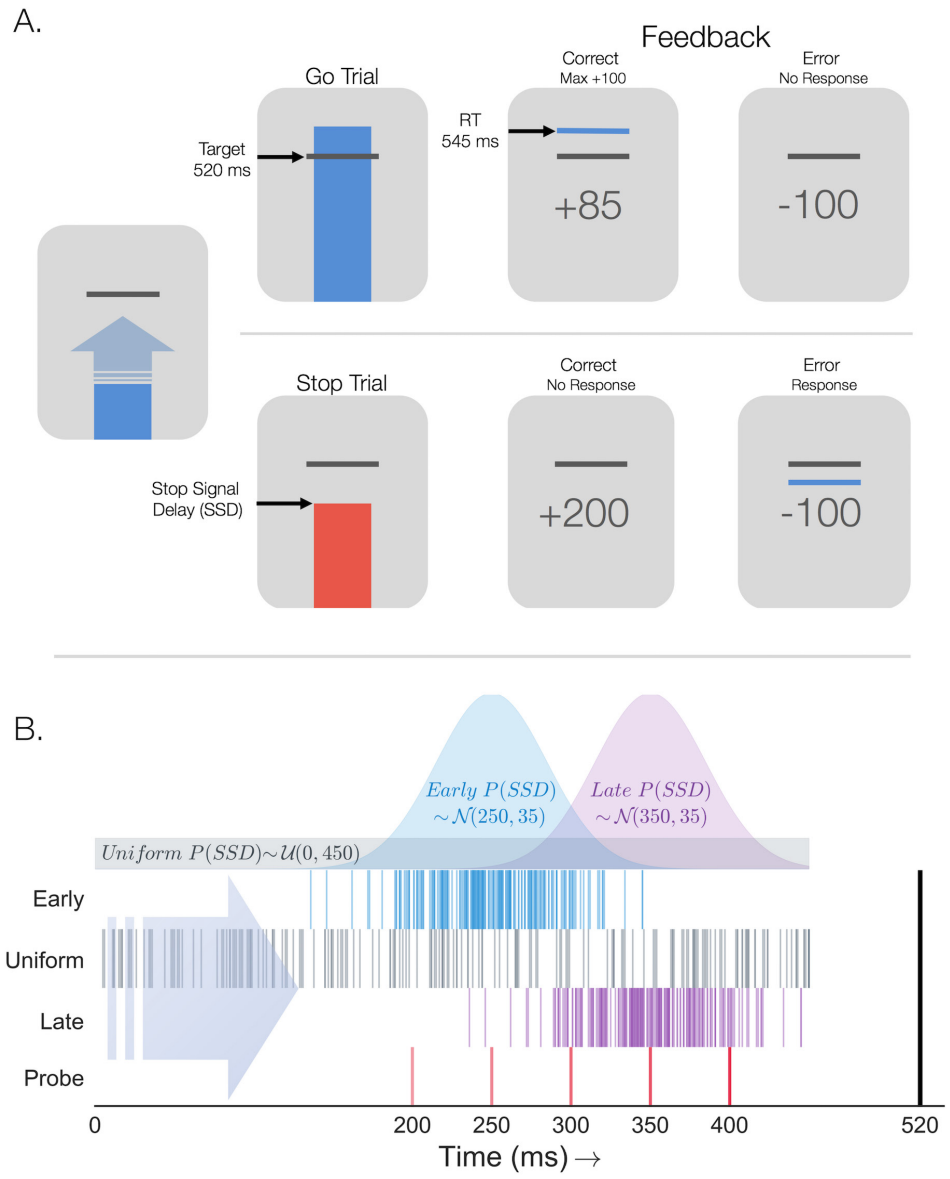

Figure 8. Adaptive Stop-Signal Task and Contextual SSD Statistics

(**A**) Behavioral stop-signal experiment. On Go trials (upper) subjects are instructed to press a key when the moving bar crosses a target line, which always occurs on 520ms after trial onset. Feedback is given informing the subject if their response was earlier or later than the Go Target (max +100 points). On Stop trials (lower), the bar stops and turns red prior to reaching the Target line. If no response is made (correct), the subject receives a bonus of +200 points. Failure to inhibit the keypress results in a -100 point penalty. (**B**) Stop-Signal statistics across Contexts. Distributions show the sampling distributions for SSDs on Context trials in the Early (blue), Uniform (gray), and Late (purple) groups. Early and Late SSDs were Normally distributed (parameters stated as in-figure text $\mathcal{N}$~(μ, σ)). Below the distributions, each row of tick-marks shows the Context SSDs for a single example subject each group. Bottom row of red tick-marks shows the five Probe SSDs included for all subjects regardless of Context.

3.2.3 Static Dependent Process Model (DPM)

The dependent-race model (DPM) assumes that action-facilitating (i.e., direct) and action-suppressing (i.e., indirect) signals are integrated over time as a single execution process (θ_e), with a drift-rate that increases with the ratio of direct-to-indirect pathway activation. The linear drift and diffusion (φ_e) of the execution process is described by the stochastic differential equation in Equation 1, accumulating with a mean rate of v_e (i.e., drift rate) and a standard deviation described by the Wiener diffusion process (e.g., white noise) with diffusion constant σ. The execution process is fully described by Equation 2 in which the linear accumulation described by Equation 1 is scaled by an urgency signal, modeled as a hyperbolic cosine function of time with gain γ.

$$d\varphi_e = \upsilon_e dt + \sigma dW \quad (1)$$

$$\theta_e(t) = \varphi_e(t) \cdot cosh(\gamma \cdot t) \quad (2)$$

A response is recorded if θ_e reaches the execution boundary (a) before the end of the trial window (680ms) and before the braking process reaches the lower (0) boundary (see below). In the event of a stop cue, the braking process (θ_b) is initiated at the current state of θ_e with a negative drift rate ($-v_b$). If θ_b reaches the 0 boundary before θ_e reaches the execution boundary then no response or RT is recorded from the model. The in θ_b over time is given by Equation 3, expressing the same temporal dynamics of θ_e but with a negative drift rate ($-v_b$) and in the absence of the dynamic bias signal. The dependency between θ_b and θ_e in the DPM is described by the conditional statement in Equation 4, declaring that the initial state of θ_b (occurring at t = SSD) is equal to the state of θ_e(SSD).

$$d\theta_b = v_b dt + \sigma dW \quad (3)$$

$$\theta_b(SSD) = \theta_e(SSD) \quad (4)$$

The DPM was fit to the average stop accuracy at each Probe SSD and the correct and error RT distributions (10^{th}, 20^{th}, 30^{th}… 90^{th} quantiles) in each Context group using a combination of global and local optimization techniques (Bogacz & Cohen, 2004; Dunovan et al., 2015). All fits were initialized from multiple starting values in steps to avoid biasing model selection to unfair advantages in the initial settings. Given a set of initial parameter values, all model parameters – Execution Drift-Rate (v_e), Brake Drift-Rate (v_b), Execution onset delay (tr), boundary height (a) and dynamic gain (γ) were optimized by minimizing a weighted cost function (see Equation 5 below) equal to the summed and squared error between an observed and simulated (denoted by $\wedge$ hat symbol) vector containing the following statistics: probability (P) of responding on Go trials (g), probability of stopping at each Probe SSD (d={200, 250, 300, 350, 400ms}), and RT quantiles (q={.1, .2, …, .9}) on correct (RT^C) and error (RT^E) trials.

The cost-function weights (w) were derived by first taking the variance in each summary measure included in the observed vector (across subjects), then dividing the mean variance by the full vector of variance scores. This approach represents the variability of each value in the vector as a ratio (Ratcliff & Tuerlinckx, 2002), where values closer to the mean are assigned a weight close to 1, and values associated with higher variability a weight <1, lower variability a weight >1 (Bogacz et al., 2006; Dunovan et al., 2015). Weights applied to the RT quantiles were calculated by estimating the variance for each of the RT quantiles (Maritz & Jarrett, 1978) and then dividing the mean variance by that of each quantile. Stop accuracy weights were calculated by taking the variance in stop accuracy at each Probe SSD (across subjects) and then dividing the mean variance by that of each condition.

$$\chi^2_{static} = w_g \cdot (P_g - \hat{P}_g)^2 + \sum_{d}^{5} w_d \cdot (P_d - \hat{P}_d)^2 + \sum_{q}^{9} w_q^C \cdot (RT_q^C - \hat{RT}_q^C)^2 + \sum_{q}^{9} w_q^E \cdot (RT_q^E - \hat{RT}_q^E)^2 \quad (5)$$

In order to get an estimate of fit reliability for each model I restarted the fitting procedure from 20 randomly sampled sets of initial parameter values. Each initial set was then optimized to average data in the Uniform condition using the Basinhopping algorithm (Wales & Doye, 1997) to find the region of global minimum followed by a Nelder-Mead simplex optimization (Nelder & Mead, 1964) for fine tuning globally optimized parameter values. The simplex-optimized parameter estimates were then held constant except for one or two designated context-dependent parameter(s) that were submitted to a second Simplex run in order to find the best fitting values in the Early and Late conditions.

3.3 RESULTS

3.3.1 Behavior

To assess the behavioral differences across Contexts, I compared accuracy on stop-signal trials at each Probe SSD (200, 250, … 400ms) across groups as well as the mean RTs on correct (response on Go trials) and error (i.e., response on Stop trial) responses. Separate one-way ANOVAs revealed a significant effect of Context across groups on both correct RTs, $F(2,72) = 10.07$, $p=.00014$, and error RT (stop-signal respond trials), $F(2,72) = 21.722$, $p<.00001$. A mixed-effects ANOVA was run to determine the statistical significance of differences in mean

stop accuracy across probe SSDs (within-subjects) and SSD Context (between-subjects). Reported statistics are corrected for sphericity violation using the Greenhouse-Geisser method (Abdi, 2010). Consistent with our hypothesis, I found a significant interaction between Context and Probe SSD, $F(2.226,80.152)=3.604$, $p=.027$, supporting our hypothesis that reactive stopping ability varies as a function of experience and expectations about control demands over time. Specifically, frequent exposure to later stop-signals led to delayed responding on Go trials as well as greater stopping success across Probe trial SSD's, exhibited by the rightward shift in the stop-curve and RT distributions in the Uniform and Late conditions (see Figure 9).

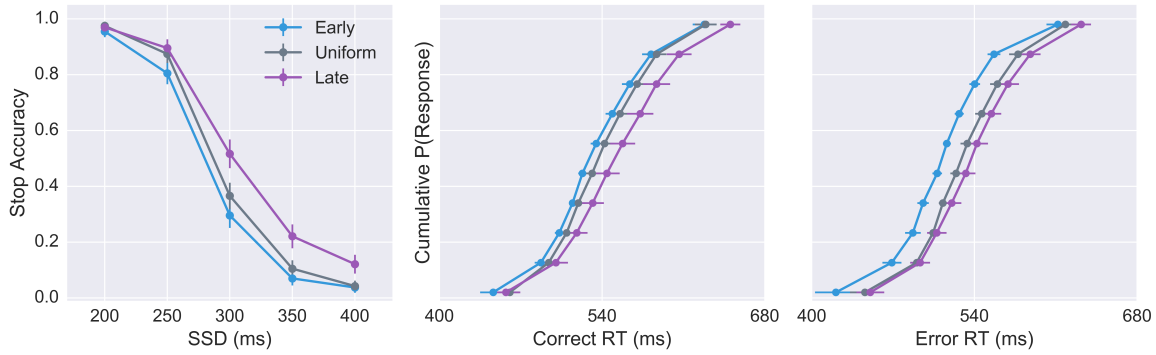

Figure 9. Effects of Context on Stop Accuracy and Response Times

Subject-averaged stop-accuracy (left) and cumulative RT distributions for correct (Go trial; middle) and error (Stop trial; right) responses for all three Context conditions. Error bars reflect the 95% confidence interval (CI) calculated across subjects.

3.3.2 Static Model Fits

In order to determine which of the model parameter(s) best accounted for the observed behavioral effects across Contexts, I first fit the model to the average data in the Uniform group, leaving all parameters free (see Table 1). Using the optimized Uniform parameter estimates to initialize the model, I then fit different versions of the model to data in the Early and Late groups

allowing only one or two select parameters to vary between conditions. This form of nested model comparison provides a straight-forward means of testing alternative hypotheses about the mechanism underlying Context-specific adaptation.

Table 1. Average Uniform Parameters and Static Model Fit Statistics

	a	v_E	v_B	*tr*	γ	χ^2	AIC	BIC
Mean	.629	1.146	-1.012	.0746	1.149	.013	-173.19	-178.61
2xSEM	.0221	.0651	.0464	.0086	.2739	8e-4	1.064	1.064

For each model, the final Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were calculated based on the fit to both Early and Late conditions (Figure 10A). These measures reflect the goodness-of-fit for alternative models while also taking into account the complexity of each model by placing a penalty on the number for free parameters. A difference of 7-10 in the IC values for two models provides strong support for the model with the lower value. The model fits supported our previous findings, showing that contextual changes in control measures are best captured by modulation of the execution drift-rate. In Figure 10C, simulated data from the drift model are overlaid on the observed stop-accuracy curve and RT distributions in each of the groups, showing a high degree of precision in the model's prediction across measures in each group. In addition to outperforming alternative models, it is worth juxtaposing the model's goodness-of-fit with its relatively parsimonious assumptions – accounting for complex group differences in the full stop accuracy curve shape as well as correct and error RT distributions with only three degrees of freedom.

I first compared alternative models in which only one parameter was free to vary across context (see Table 2) - either execution drift, braking drift, urgency, or boundary height. In line

with our previous results, leaving the execution drift-rate free provided the best account of stop-accuracy and RT differences across contexts (Figure 10A; AICv_e =-355.55). The next best fit was afforded by allowing the urgency to vary across conditions (|AICv_e–AICγ| = 7.65). To further test the relationship between execution drift and Context, I performed another round of fits to test for possible interactions between the execution drift-rate and a second free parameter (a, v_b, γ).

Table 2. Static Fit Statistics for Early and Late Contexts

Context Parameter	χ^2	AIC	BIC
Execution Drift (v_E)	.019	-355.55	-354.6
Boundary Height (a)	.039	-335.14	-334.58
Braking Drift (v_B)	.038	-336.64	-336.07
Urgency Gain (γ)	.030	-347.90	-347.34
a & v_E	.0236	-355.88	-358.63
v_B & v_E	.0260	-349.84	-352.58
γ & v_E	.0266	-349.03	-351.77

The AIC and BIC scores from these fits showed that a combination of execution drift and boundary height provided a slightly better fit than execution drift alone (Figure 10A; see Table 2). In Figure 10B the fitted drift-rate and boundary height estimates are plotted, showing the change in value between Early and Late condition. It can be seen that, when drift-rate is fixed, the boundary height parameter shows a significant increase in the Late relative to the Early condition - accounting for the slower response times and higher stop accuracy. This effect on boundary height becomes diminished however in the model that also both parameters to vary across conditions, whereas the change in drift-rate remains largely the same when additional free parameters are included (black line shows the average of effect across all dual-parameter models). This asymmetry in the relative effect sizes suggests that the drift-rate is the primary

driver behind the observed behavioral effects and that the small improvements in fit quality with both drift-rate and boundary height free is not strong evidence for a dual mechanism account. Thus, given the parsimony and high quality fits of the drift-only model, I will focus on this hypothesis in all subsequent analyses of this chapter.

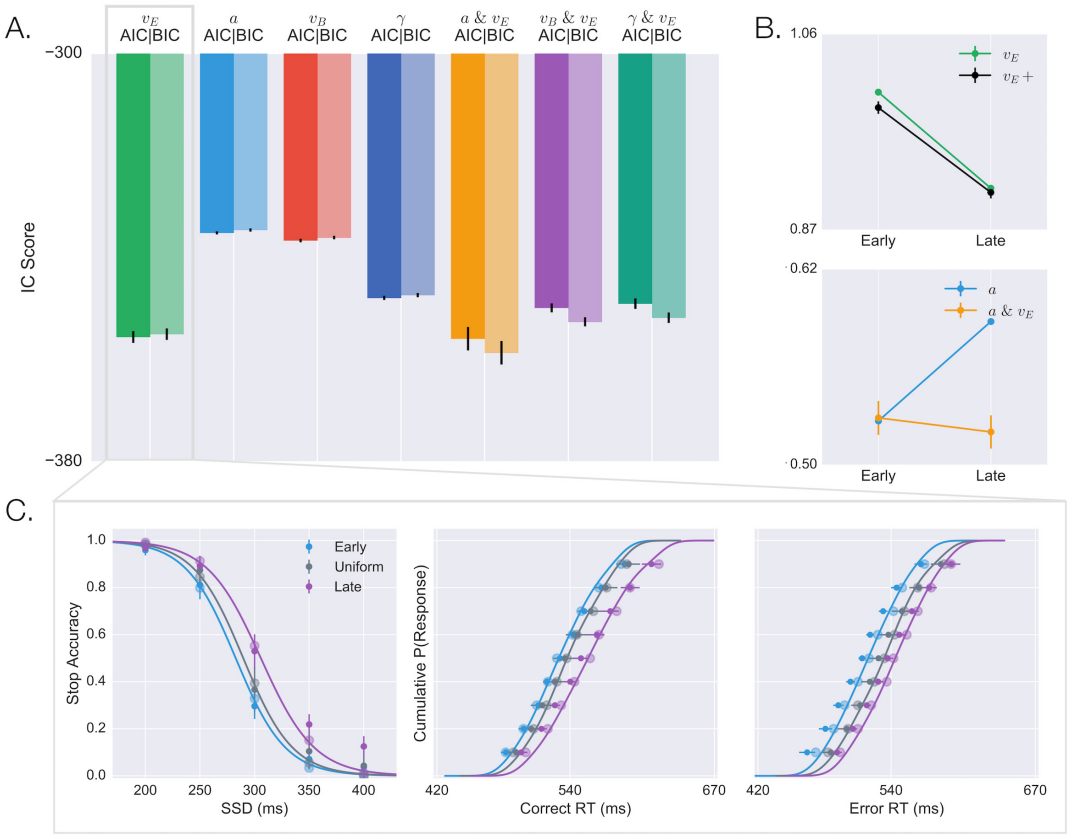

Figure 10. Model Comparison and Best-Fit Predictions Across Context

(**A**) AIC (dark) and BIC (light) scores for all single-parameter models, allowing either execution drift-rate (v_e; green), boundary height (a; cyan), braking drift-rate (v_b; red), or urgency (γ; blue) to vary across Context conditions., Three dual-parameter models were also included to test for possible benefits of allowing v_e (best-fitting single parameter model) to vary along with either a (yellow), v_b (purple), or γ (teal). (**B**) Average drift-rate parameter estimates (top panel) in the Early and Late Contexts from the best-fit result in the single-parameter model (green) plotted alongside the corresponding change in drift-rate averaged across all dual-parameter models (black). Same comparison is shown for the boundary height parameter (bottom panel), which exhibits a reduction in effect size when accompanied by a free drift-rate (yellow) compared to when the drift-rate is fixed (cyan). (**C**) Average model fits (lines and larger transparent circles) overlaid on the average empirical data for Early (blue), Uniform (gray), and Late (purple) Contexts.

3.3.3 Adaptive Model Fits

In most decision-making paradigms choices are repeatedly made between stimuli that vary along one or two dimensions of interest (signal-to-noise, frequency, etc.) but without any serial dependence between trials. Even when contextual factors are manipulated, such as the relative probability or value of different outcomes, these contingencies are usually conveyed to the subject explicitly or, in the case of animal research, overlearned prior to data collection. In this task, however, the behavioral differences across groups are driven by trialwise adaptation to the specific control demands of each Context. The previous model fits showed that modulation of the execution drift-rate best account for the aggregate effect Early, Uniform and Late Context trial distributions on behavioral performance. However, it is not clear from this analysis whether error-driven changes in drift-rate are able to capture trial-to-trial adjustment of RT and stop accuracy as statistics of the environment are learned experientially. In the next section I show how simple feedback-dependent tuning of the drift-rate gives rise to observed behavior, improving fits to the average data while also accounting for differences in the behavioral timecourse of learning across Contexts.

On correct Go trials, the drift-rate is updated to reflect the signed difference between simulated RT on the current trial (RT_t) and the Target time (T^G= 520ms). Figure 11A shows an example of how the learning rule expressed by Equation 6 alters the drift-rate in response to timing errors on 'Go' trials (Figure 11A, left), increasing the drift-rate following "slow" responses (>520ms, Figure 11A, upper-left) and decreasing the drift-rate following "fast" responses (<520ms, lower-left).

$$v_{t+1} = v_t + \alpha \cdot (v_t - v_t \cdot e^{[T - RTt]}) \tag{6}$$

On Stop trials (Figure 11A, right), the drift-rate is updated by the same rule as on Go trials but instead of reflecting the temporal difference between the RT and T^G, the temporal error in RT is calculated with respect to T^S. The T^S parameter is initialized to the time-boundary of the trial ($T^S(0)$ = 680ms) and controls the magnitude of drift-rate suppression on failed "Stop" trials. Conceptually, T^S serves a similar function as T^G, acting as a temporal anchor that can be compared with the RT on each trial to obtain an estimate for how severely to penalize incorrect (as opposed to correct) "go" responses. In contrast to T^G, which is a constant value (both across trials and Contexts), T^S is a latent parameter estimated by the observer based on their expectations about the probability of a stop-signal across time. For instance, on the first failed Stop trial, the RT error (T^S_t-RT_t) causes the drift-rate to be suppressed enough to decrease the probability of the execution process reaching the boundary before 680ms. However, as the stop-signal statistics are learned in each Context, the magnitude of drift-rate suppression following failed stops should diminish over time, maximally in the Early context (i.e., when late SSDs are unlikely) and minimally in the Late Context (i.e., when late SSDs are more likely). To account for this, T^S was decreased following successful 'Stop' trials (Figure 11A, upper-right) and increased following trials when the model incorrectly executed a response (Figure 11A, lower-right). The trial-outcome contingencies and learning-rate (β) used to update T^S are expressed by Equation 7 below.

$$T^S_{t+1} = \begin{cases} T^S_t + \beta \cdot T^S_t, & \text{if Stop Failure} \\ T^S_t - \beta \cdot T^S_t, & \text{if Stop Success} \end{cases} \quad (7)$$

To obtain estimates for the learning rate parameters for adaptation in the drift-rate (α) and T^S (β), this adaptive form of the model was re-fit to the RT and stop accuracy data in the Uniform condition, holding all previously estimated parameters constant. Because standard

parameter optimization for accumulator models requires information about the variance of response-times across trials, these approaches are poorly suited for investigating how decision parameters respond to error on a trialwise basis. To overcome this issue, the cost function was modified to identify the values for α and β that minimized the difference between the average observed and model-predicted RT and stop accuracy over a moving window of about 30 trials (Equation 8). By averaging the behavior in 30-trial bins (30 bins total), this ensured that multiple Stop trials were included in each bin while still allowing relatively high-frequency behavioral changes to be expressed in the cost function. Also, these fits were performed by iteratively simulating the same trial sequence as observed for each individual subject, and fitting the average simulated subject to the average observed subject. This ensures that direct comparisons can be made between the trajectory of learning in the model and actual behavior.

$$\chi^2_{adapt} = \sum_{i}^{30}(\mu_{acc,i} - \hat{\mu}_{acc,i})^2 + \sum_{i}^{30}(\mu_{rt,i} - \hat{\mu}_{rt,i})^2 \quad (8)$$

In Figure 11B, the decision outcome ("Go" green dots; "Stop", red dots) and Go RTs are shown for 120 trials of the experiment for a single representative subject, with shaded boxes illustrating the bin width used to calculate the timecourse of average RT (top) and stop-accuracy data (bottom) displayed in Figure 11C. In Figure 11C, the model-predicted change in Go RT and Stop accuracy across the experiment is overlaid on the corresponding empirical measure, demonstrating a high degree of overlap between the two. To confirm that the trial-averaged behavior of the model was preserved after fitting the learning rates, the stop-accuracy curve and RT statistics were calculated from simulations of the adaptive model and overlaid on the average Uniform data (Figure 11D). The model's predictions are indeed closely aligned with all empirical statistics (adaptive χ^2=.0042, static χ^2 = .01). While this is not necessarily surprising, it

is promising to see that introducing feedback-dependent adaptation in the drift-rate leads to improved fits to the average data despite not including these statistics in the learning cost function.

After confirming that error-driven changes in the drift-rate were able to account for the temporal changes in behavior observed in the Uniform Context, I next asked if this mechanism was sufficient to account for observed differences in the Early and Late Contexts. If so, then these differences should emerge when the adaptive model is tasked with balancing the same goals of maximizing RT precision and minimizing inhibitory failures against environments with lighter/heavier control demands. In other words, the adaptive model should recover the observed behavioral profiles of the Early and Late Context groups when exposed to the same trial-structure as the subjects experienced. To test this prediction, I initialized the adaptive model with the best-fitting Uniform parameters (including the learning rates as in Figure 11C) and ran simulations of the Early and Late conditions using the same trial sequence as observed by the subjects in each Context.

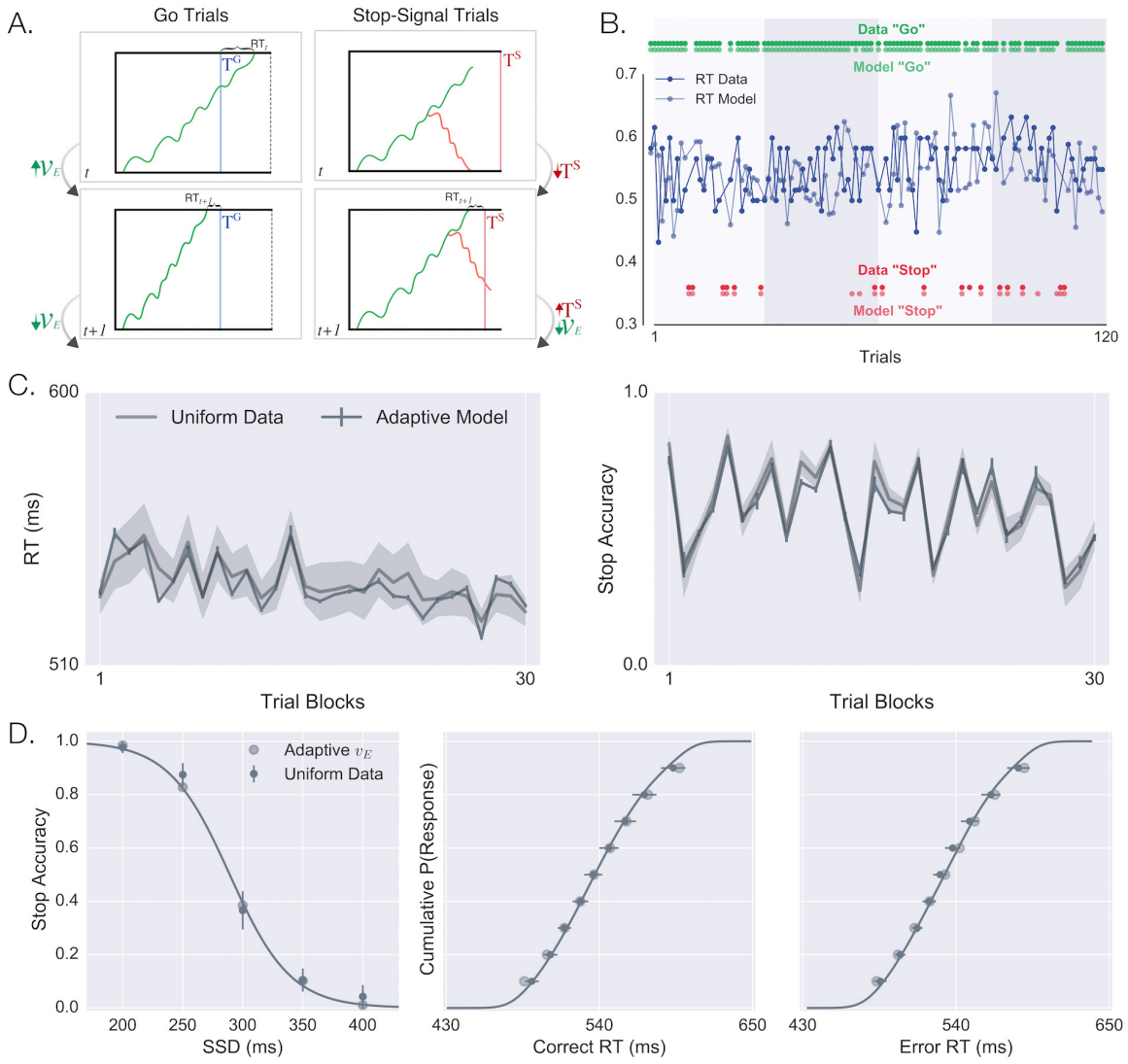

Figure 11. Adaptive DPM and Predicted Learning Trajectory in Uniform Condition

(**A**) Schematic showing how the execution drift-rate is modulated following timing errors on Go trials (left) and how T^s is modulated following successful and failed inhibitions on Stop trials. (**B**) Raw data from a single subject (dark colors) and the model's (light colors) performance on the same 120 trials. Go-trial RT's are shown in blue as a timeseries, "Go" and "stop" decisions are shown as green and red dots at the top and bottom of the plot, respectively. Light and dark shaded areas display the bin width over which RT and stop-accuracy data were averaged during model-fits (**C**) Subject-averaged timeseries (dark line) and 95%CI (transparent gray) showing the RT on Go trials (left) and accuracy percentage on Stop trials (right). Both timeseries are 30 points in total, each calculated as by taking the average RT/stop-accuracy over successive~30-trial windows). The corresponding model predictions are overlaid (light gray line), averaged over simulations to each individual subject's data. (**D**) Average empirical stop-accuracy and RT statistics in the Uniform condition, (same as shown in Figure 9) with the predictions generated from simulations with the adaptive DPM.

Figure 12A shows the simulated stop-curve and RT distributions generated by the adaptive DPM based on feedback in the Early and Late Contexts. As in the observed data, adaptation to Early SSDs led to impaired stopping accuracy but faster RT's relative to simulated predictions in the Late Context. In Figure 12B-C, the middle panels show the same trial-binned RT and stop-accuracy means as in Figure 11C (Uniform Context), flanked by corresponding timecourses from simulations to Early (left) and Late (right) Contexts. The adaptive model predictions show a high degree of flexibility, conforming to idiosyncratic changes in the trialwise dynamics in behavior across Contexts. For instance, the RTs in the Early Context exhibit a relatively minor but gradual decay over the course of the experiment (Figure 12B, left), contrasting markedly from the sharp early increase and general volatility of RTs in the Late Context (Figure 12B, right). The adaptive model largely captures both patterns, underscoring feedback-driven adaptation in the drift-rate as a powerful and flexible tool for commanding inhibitory control across a variety of settings. In addition to predicting group differences in the timecourse of RTs, the simulations in Figure 12C show a striking degree of precision in the model-estimated changes in stop-accuracy, both over time and between groups.

Finally, Figure 13A shows the adaptive drift-rate in each Context across the same trial blocks as shown in Figure 12C. The relative variability in drift-rate adaptation across all trials in each Context is summarized by the box-and-whisker plots in Figure 13B, with parameter estimates from the static model overlaid, showing that adaptive changes in drift-rate have a central tendency equal to the value that best describes the trial-averaged behavior in each group. Collectively, the results of the adaptive simulations suggest that goal-directed tuning of movement timing (RT) and control (stopping efficacy) can be achieved via feedback-dependent learning in the drift-rate.

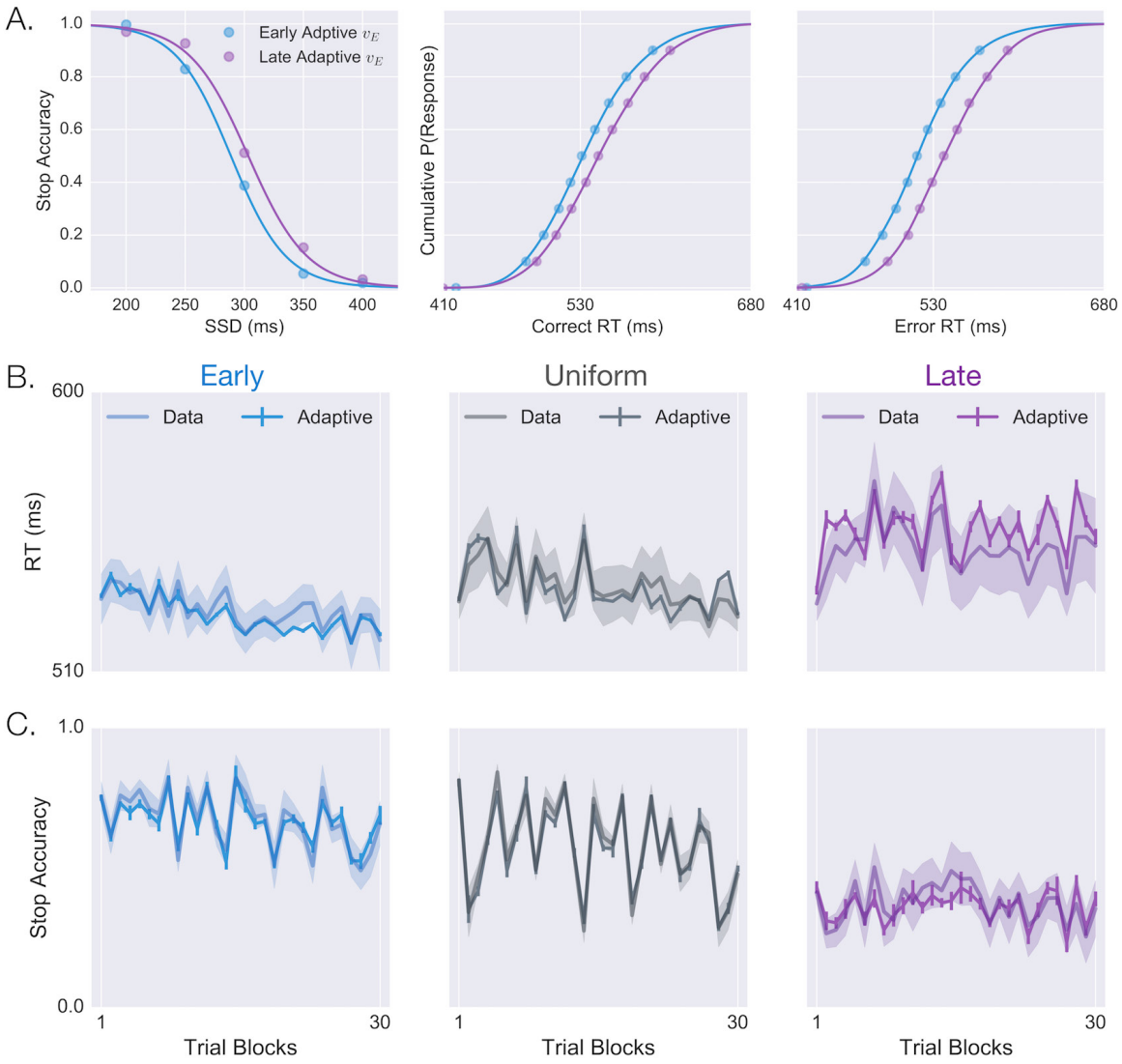

Figure 12. Adaptive DPM Modulates Behavior to Context-Specific Control Demands

(A) Adaptive DPM simulations based on trial sequence and SSD's in the Early (blue) and Late (purple) Contexts, showing the model's trial-averaged stop-accuracy curve (left) and cumulative RT distribution on correct (Go) and error (Stop) trials. **(B)** Empirical timeseries of Go RT's with model predictions overlaid (calculated using the same method described for Figure 11C) for Early (left), Uniform (middle, same as in Figure 11C), and Late (right) Contexts. **(C)** Empirical and model predicted timeseries of stop-accuracy for the same conditions as in panel B.

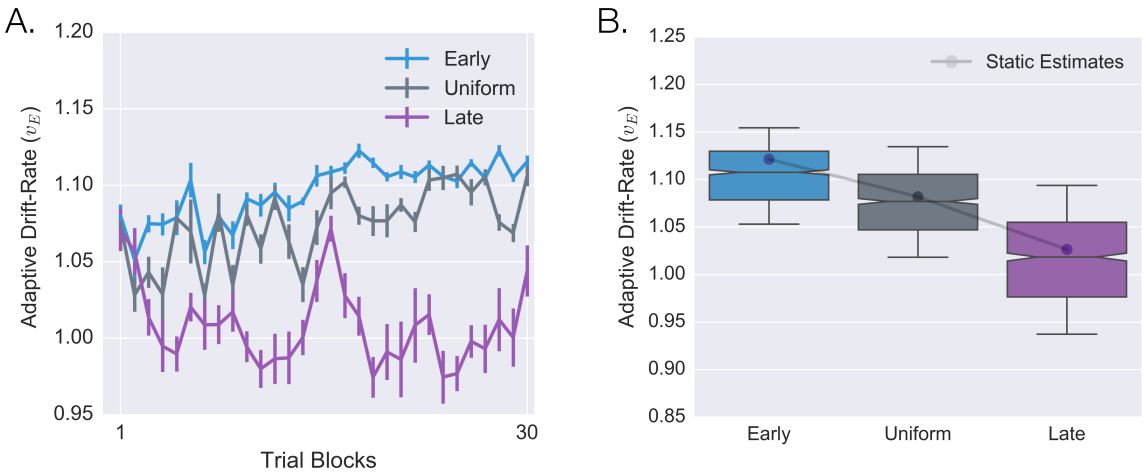

Figure 13. Drift-Rate Adaptation to Feedback Recovers Static Model Parameters

(**A**) Timeseries plots showing the change in adaptive execution drift-rate (v_E) across trial blocks (same bin width as RT and accuracy timeseries shown in Figure 12B) from the Adaptive DPM simulations in the Early (blue), Uniform (gray), and Late (purple) Contexts. Error bars reflect the 95%CI, calculated across subjects. All models were initialized with the optimized parameter estimates for the Uniform Context, allowing only the execution drift-rate to vary across trials. (**B**) Boxplots show the distribution of trialwise estimates of v_E from the Adaptive simulations for the average subject in each Context (same color conventions as in A). The dot markers overlaid on each boxplot represent the optimized drift-rate values in the Static DPM. Boxplot whiskers depict the 1.5 s.d. around the mean.

3.4 SUMMARY OF RESULTS

In contrast to most theories that assume a sensory-driven and cortically implemented drift-rate, the DPM posits that the drift-rate is at least partially a reflection of competing signals of the direct and indirect pathways in the BG. One critical prediction that follows from this assumption is that feedback about decision outcomes should modulate the drift-rate in ways that align behavioral control with the demands of the environment and facilitate goal acquisition.

Particularly relevant to the current study, recent work has found that the weighting of cortico-striatal synapses on direct and indirect MSNs can be incrementally altered to generate faster or slower movements, and that this effect disappears in the presence of dopamine

antagonist (Yttri & Dudman, 2016). The authors found that closed-loop stimulation of the direct pathway increased movement velocity when stimulation was selectively yoked to faster movements and decreased movement velocity when yoked to slower movements. Surprisingly, the exact opposite effect was observed when stimulation was applied to indirect pathway neurons, increasing velocity when stimulation was applied to slower movements and vice versa. In other words, activity-dependent plasticity in both pathways was sufficient to bi-directionally alter task-relevant movement kinematics, revealing an unprecedented degree of control by BG circuitry. This study provides a useful computational framework for studying the effects of reward feedback on cortico-striatal encoding of action timing.

4.0 ADAPTIVE DECISIONS BETWEEN MULTIPLE ALTERNATIVES

In the preceding section (see 3.0), I examined how the underlying dynamics of single-action decisions change with experience, showing how feedback dependent plasticity in the execution drift-rate incorporates control demands of the environment with task goals to drive adaptive behavior. By allowing the drift-rate parameter to change in response to feedback, the adaptive DPM was able to capture the same trial-wise fluctuations in RT and stop-accuracy as observed experimentally in each Context condition. The gradual tuning of the execution process is consistent with the well-studied effects of dopaminergic reinforcement on the direct and indirect pathways as (see 2.3). Moreover, these findings are in line with recent work in human neuroimaging (van Maanen et al., 2016) and optogenetic studies in mice (Panigrahi et al., 2015; Yttri & Dudman, 2016) showing a clear functional link between the striatum and the kinetics of goal-directed movements (see Dudman & Krakauer, (2016) for a current review). Thus, rather than simply opening the gate for action execution, cortico-striatal pathways appear to play an important role in shaping *how* actions are expressed so as to minimize errors in specific dimensions. This is an important distinction, as current RL models of BG computations do not offer testable predictions about the relevant environmental factors, cognitive mechanisms, or neural implementation of movement parameters such as the timing or velocity of action execution.

As a result, the experimental paradigms that have been developed to study RL phenomena have largely relied on single discrete measures of decision outcome. One such paradigm is the IGT (Bechara, Damasio, Damasio, & Anderson, 1994), widely considered to be the gold-standard behavioral task for studying value-based decision-making. In the IGT, the subject makes sequential draws from four decks of cards and receives feedback in the form of rewards and penalties. Each deck has a feedback schedule that varies along two dimensions of interest (see Table 3): 1) long-term payoff and 2) the frequency of positive versus negative feedback.

Table 3. Target feedback schedules modeled after decks in the Iowa Gambling Task

	Low Value Low Frequency	Low Value High Frequency	High Value Low Frequency	High Value High Frequency
Target ("Deck")	A (Blue)	B (Red)	C (Purple)	D (Green)
Reward (+)	+100	+100	+50	+50
Penalty (-)	-100 to -350	-1250	-50	-250
+/- Ratio	5:5	9:1	5:5	9:1
Net Sum	-250	-250	250	250

With respect to long-term payoff, decks A and B are considered "bad" decks as draws from these decks have a negative cumulative value and lead to long-term loss. In contrast, decks C and D have a positive cumulative value and lead to long-term gains, thus drawing from these decks is considered an optimal strategy over drawing from decks A and B. The second dimension of interest, reward frequency, is assessed by comparing the number of times the subject draws from decks A and C, which both yield positive rewards with 5 to 1 odds, with the number of times the subject draws from decks B and D, which yield rewards at 9 to 1 odds. The standard approach to

evaluating performance on the IGT involves computing scores that reflect a subject's sensitivity to Payoff (P): the number of draws from decks C and D minus the number of draws from A and B; and Sensitivity (Q): the number of draws from decks B and D minus the number of draws from A and C.

The original prediction from the IGT is that human participants are largely optimal in learning to draw from "good" decks that maximize their long-term gains and to avoid decks that increase the risk of long-term losses (Bechara et al., 1994). As it turns out, this conclusion is often not correct, with some studies suggesting that healthy adults regularly exhibit sub-optimal decision strategies that focus on gain-loss frequency at the expense of long-term gains (Horstmann, Villringer, & Neumann, 2012). Typically, this suboptimal strategy manifests in persistent draws from deck B - termed the 'prominent deck B phenomenon' (Lin, Chiu, Lee, & Hsieh, 2007)- as well as fewer draws from Deck C - termed the 'sunken deck C phenomenon' (Chiu & Lin, 2007). However, because of the way P and Q are calculated, many studies have overlooked this strategy as a sub-optimal, yet normative, component of behavior in healthy adults. The consequence of this oversight bears on the validity of performance deficits in the IGT that have been widely used as confirmatory evidence of impaired judgment in individuals with ADHD (Toplak, Jain, & Tannock, 2005) and schizophrenia (Shurman, Horan, & Nuechterlein, 2005). Furthermore, the standard approach of collapsing across deck selections may explain why, despite the hundreds of studies that have confirmed the hypothesis of normative gain maximization (Hawthorne & Pierce, 2015), computational models of the IGT widely disagree with each other on the nature of adaptive, value-based decision making (Steingroever, Wetzels, & Wagenmakers, 2013; Worthy, Hawthorne, & Otto, 2013). One potential reason for the discontinuity between strictly behavioral studies of the IGT and those that attempt to model the

behavior computationally is that choice outcomes are not sufficient indicators of the underlying cognitive mechanisms. This argument is underscored by considering that models of binary perceptual decisions, arguably a simpler form decision-making than that probed by the IGT, would be impossible to distinguish without access to both decision outcomes and RTs.

In this section, I investigate behavioral performance on a novel sensorimotor variant of the IGT where, instead of selecting virtual cards from a deck, human participants perform a center-out reach task to peripheral targets on a screen, each mirroring the feedback schedule of a deck in the original task. The primary motivation behind this design is to expand the behavioral profile of risk- and value-based decision-making by measuring the temporal and kinematic effects of reward frequency and long-term gains. In addition, I can control the degree of certainty in choice selection by increasing the spatial variance of the targets, which as I pointed out in Chapter I should push individuals to become more exploratory in their decisions, rather than exploitative. In order to theoretically motivate the expected results, I will use a simulation-based hypothesis generation approach that leverages knowledge about the involvement of cortico-BG circuitry in regulating these phenomena to generate specific predictions about the underlying causes of suboptimal policies and other standing debates in the literature.

In section 3.2.3, I took the standard approach of optimizing decision and learning-rate parameters of the DPM to accuracy and RT data. Here I take a different approach, using a simplified network model to generate and test predictions about dimensions of interest in the data. There are a several advantages to prediction-oriented approaches that address specific weaknesses of model-fitting (Thura, 2016). One of the greatest strengths of accumulator models is that they offer a highly flexible and assumption-free mechanism for describing any bounded stochastic system with analog input and bounded digital output. However, greater flexibility

often comes at a cost to model identity (Palmeri, Schall, & Logan, 2013), the measure of how uniquely predictive a model is of the phenomenon it intends to describe. Optimization algorithms can exploit this flexibility and lead to convincing visual fits to the data that provide spurious evidence for the predictive utility of a model that is in fact, highly general. One way to avoid a model identity crisis is to develop behavioral paradigms that force models to disagree and therefore be falsified. Another way is to leverage previous observations to generate conditional predictions about how the system should behave in a novel setting. In the present study, I have two important sources of prior knowledge: 1) the literature reviewed in section 2.0 regarding the channel-like architecture of the BG and the physiological processes that drive learning and action-selection; and 2) the observed effects of feedback on reaching behavior in the current task (see 4.2.1), which provide important constraints for setting up and testing novel predictions about untested dimensions of behavior.

4.1 METHODS

4.1.1 Participants

Neurologically healthy adults (N=35, Mean age 28 years) were recruited from greater Pittsburgh area through a community subject pool sponsored by Carnegie Mellon University. All procedures were approved by the local Institutional Review Board. Two subjects were excluded due to a programming error that resulted in erroneous tracking of the reach trajectory and timing. All subjects were informed that they were guaranteed $10 compensation for participating with the chance to earn up to $5 bonus based on their task performance. However, this incentive was

purely motivational and all subjects were compensated the full $15 for their time. The task was self-paced but typically lasted no more than 1.5 hours, including time dedicated to consenting, instruction, and practice. All subjects were given the opportunity to practice the task until they felt comfortable with the reaching equipment and general goals of the task.

4.1.2 Reaching Task

Subjects completed a center-out reach task using a stylus and pad to reach toward one of four spatially presented targets (2D-Normal distribution of dots) with feedback schedules modeled after those of the IGT. Subjects were situated in a 2-D virtual reaching environment and movement trajectories were recorded using a Wacom Intuos Pro Large tablet. Visual feedback was presented on a computer monitor perpendicular to the tablet workspace. Visual presentation of the hand position (filled circle) and targets were aligned so as to be veridical with the workspace of the tablet. At the beginning of each trial, subjects moved their right hand to the center of the workspace, indicated by an unfilled white circle. Once in the start circle, feedback of the hand and start circle disappeared and after a random delay (500-1500ms) four targets were presented at cardinal locations (0º, 90º,180º, 270º) around the start circle. Each target was presented as a Gaussian distribution of 100 dots with the mean of each distribution located 8cm from the start circle. Targets were presented for 300ms before disappearing, at which point the subject was instructed to make a ballistic reaching movement to the center of the target location. Visual feedback of the hand position is presented at the start of the reaching movement and removed after the reach is executed. After the reach is terminated, participants are given a feedback score. This score is calculated as the product of two values: 1) distance of reach endpoint from the center of the selected target (sensorimotor estimation) and 2) a “deck” score

for that target (value estimation). This second value follows the deck score of the typical IGT paradigm where each target has an assigned value on each trial, according to a specified combination of frequency and magnitude of gains/losses (see Table 3). For each subject, Payoff or *P* was calculated by subtracting the number of reaches to low value targets from the number of reaches to high value targets: $P = (C + D) – (A+B)$. Sensitivity or *Q* was calculated by subtracting the number of reaches to targets with high reward frequency from the number of reaches to targets with low reward frequency: $Q = (B + D) – (A+C)$.

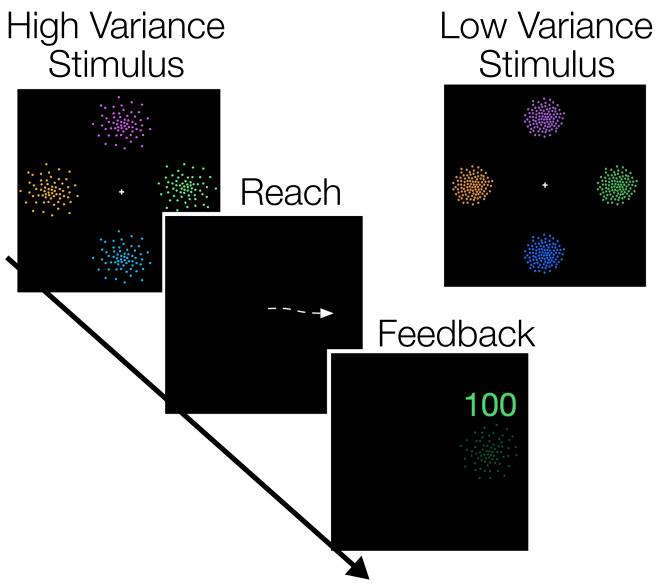

Figure 14. Multi-choice value-based reaching paradigm.

The timeline shows a single trial of a center-out reaching experiment. Subjects initiated each trial by moving stylus to the center location on a stylus, with a point-by-point mapping for visual feedback displayed on a perpendicular monitor. Once in the center, four clusters of dots were displayed following a jittered interval (see Methods), prompting the subject to make a reach selection aiming at the center of the chosen target. The reach movement was terminated as soon as movement velocity fell below a set threshold to encourage ballistic reach trajectories so as to minimize slow or overly variable movements. Once the reach was terminated a feedback score was displayed above the chosen target, with rewards shown in green and penalties shown in red font. The score reflected a weighted combination of the endpoint-error relative to the target's true center and the value returned by the target on that trial. For each of 4 blocks the subject's cumulative score was displayed in the lower left-hand corner of the screen (not shown). The upper right hand panel shows an example of the stimuli displayed in the low variance condition.

4.1.3 Adaptive Multi-Alternative Network Model

In contrast with controlling a single action (see section 3.0), the current task requires evaluation of multiple actions based on their respective histories of reinforcement. To investigate the utility of drift-rate adaptation as a mechanism for describing value-based decisions among multiple actions, I constructed a simple decision network (Figure 15) in which each target is encoded as an action channel with of independently accumulating Go (*G*) and NoGo (*N*) accumulators. For each action channel *j* and each trial *t*, the stepwise dynamics for the *G* and *N* accumulators were defined at each time step τ (Δτ = 1 ms) according to Equation 9 and Equation 10, respectively.

$$G_{j,t}(\tau) = G_{j,t}(\tau - \Delta\tau) + v_{j,t}^{G}\Delta\tau + \epsilon_{j}^{G}(\tau) \tag{9}$$

$$N_{j,t}(\tau) = N_{j,t}(\tau - \Delta\tau) + v_{j,t}^{N}\Delta\tau + \epsilon_{j}^{N}(\tau) \tag{10}$$

With v^{G} and v^{N} defining the drift rates of the G and N processes respectively. The diffusion noise on each follows a normal distribution with variance $\sigma^2 = 0.1$ (Equation 11).

$$\epsilon_{j}^{G/N} \sim \mathcal{N}(0, \sigma^2) \tag{11}$$

The execution process θ for each action channel *j* is computed in the network output layer as the difference between the *G* and *N* processes (Equation 12).

$$\theta_{j,t}(\tau) = [G_{j,t}(\tau) - N_{j,t}(\tau)] \cdot cosh(\gamma \cdot \tau) \tag{12}$$

The hyperbolic cosine term introduces a dynamic bias in the signal that approximates a collapsing decision boundary (Dunovan et al., 2015; Ratcliff & Frank, 2012) at a rate determined by the parameter γ. Each trial simulation continues until the first deck execution process reaches the decision boundary *a*. At the end of each simulated trial, the network received a feedback

signal, *r(t)*, reflecting the value of the target selected. This signal was used to update the state value of each target, x_j as shown by Equation 13.

$$x_j\,(t+1) = x_j\,(t) + \alpha \cdot [r(t) - x_j\,(t)] \tag{13}$$

On trials with positive feedback (i.e., gain), α^G is applied as the learning rate whereas on trials with negative feedback (i.e., loss), α^N is applied. Several recent computational studies have used a similar dual learning rate system to describe dissociable contributions of plasticity in the direct and indirect pathways (Cockburn et al., 2014; Collins & Frank, 2014). Here, the valence of feedback determines which learning rate is used to update the value estimate for each target, which determines the rate of learning by the "Critic" in the Actor-Critic learning model. However, in this model the "Actor" adjusts the weights of each action according to both α^G and α^N on each trial, as described below (see also Collins & Frank, (2014)). This state value function was then used to update the action selection probability p_j for each Target j given by the softmax probability function in Equation 14 (Sutton et al., 1998). Note that lowercase p_j reflects the updated probability for target *j*, as uppercase *P* is used to denote payoff throughout the text.

$$p_j(t) = \frac{e^{\beta \cdot x_j(t)}}{\sum_i^n e^{\beta \cdot x_{j_i}(t)}} \tag{14}$$

where β is the inverse temperature parameter and $x_j(t)$ is the current value estimate from Equation 13. Typically, p_j is estimated for each deck and used to perform a weighted selection from the set of possible alternatives. Here, the change in $p_j(t)$ from the previous trial $p_j(t-1)$ is calculated to obtain an estimate of the change in choice probability for each deck, $\delta_j(t)$

$$\delta(t) = p_j\,(t) - p_j\,(t-1) \tag{15}$$

This additional step effectively converts the value update from Equation 13 into proportional change in selection probability for each channel. This "choice probability" error signal is then used to update the relative drift rates of the *G* (Equation 16) and *N* (Equation 17) processes in each action channel.

$$v_{j,t}^{G} = v_{j,t-1}^{G} + \alpha^{G} \cdot \delta_t \tag{16}$$

$$v_{j,t}^{N} = v_{j,t-1}^{N} + \alpha^{N} \cdot -\delta_t \tag{17}$$

By first converting the value prediction-error into a probability, this allows for a more straight-forward update of the drift-rate term for the Go and NoGo units. Otherwise, large discrepancies in value, such as those yielded by target B, ranging between -1250 and +100, can lead to destabilizing changes in the drift-rate.

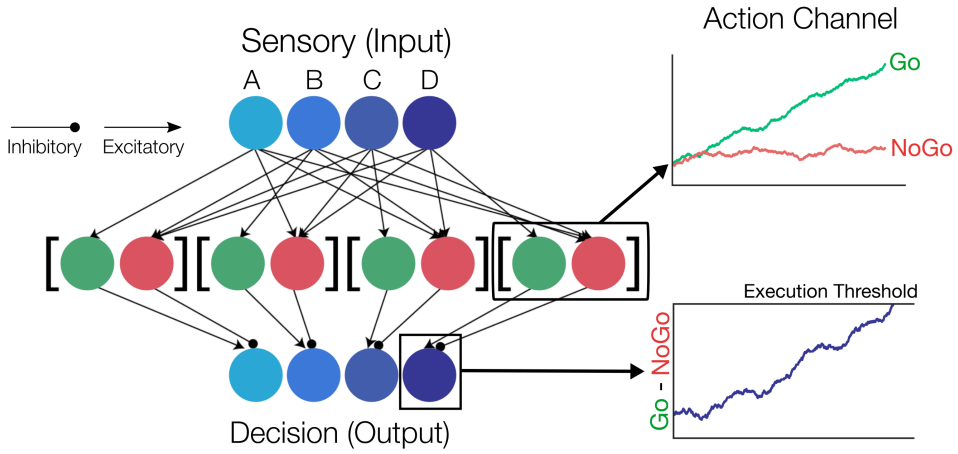

Figure 15. Network of competing action channels and parameter influence on Payoff.

The network represents each target as a separate sensory input that sends a drift-rate to the Go pathway of a single action channel and contributes to the drift-rates of NoGo pathways in the remaining action channels (i.e., center-surround architecture). Within each action channel (right inset) the Go and NoGo pathway activity accumulate independently and project to a single output decision node. The output decision node takes the difference of Go and NoGo activity, along with an urgency signal that increases with time, and accumulates to a decision boundary. The network chooses the first output node to reach its decision boundary.

4.2 RESULTS

4.2.1 Analysis of Reaching Behavior

To assess the effects of reward value and frequency, I examined the average number of reaches to each target, as well as the average reach error, response time (RT), and the duration of the reach. A one-way repeated measures ANOVA revealed a significant effect on the number of reaches, $F(3, 96)=7.862$, $p<.0001$ and a marginally significant effect on reach error $F(1, 32)=2.47$, $p=.066$. Figure 16A shows that the main effect on reach number was driven by a general preference for targets B and D over A and C. Analysis of the endpoint error of reaching movements revealed a similar pattern (Figure 16B), suggesting that subjects' target selection and subsequent reach precision were more sensitive to the frequency of rewards than overall reward magnitude. Thus, in line with previous studies of IGT performance in healthy adults, I observed evidence of both prominent deck B phenomenon (Lin et al., 2007) as well as the sunken deck C phenomenon (Chiu & Lin, 2007). Counter to my predictions, no main effect of target ID was found for RT $F(3, 96)=.852$, $p=.469$, or the duration of the reach $F(3, 96)=.852$, $p=.469$. However, inspection of the distributions of reach times (Figure 16E), showing all times for all subjects, revealed a clear difference in the positive skew of distributions for high value targets (bottom) compared to low value targets (top). One possibility is that different strategies across subjects had a wash out effect on the mean RT and duration measures. I explore this possibility in the following sections.

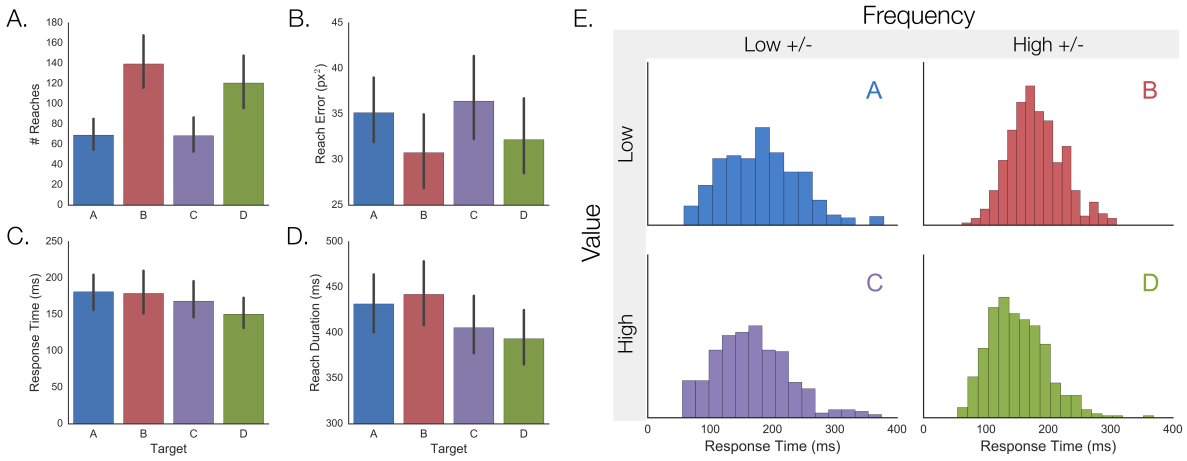

Figure 16 Summary statistics and distributions of reach durations for each target.

The subject-averaged reach count (A), reach end-point error (B), response time, calculated as the delay between stimulus offset and when the reach passed a 50-pixel radius around the center start position, and reach duration (D). The color of each bar denotes the target's feedback schedule. The target with feedback corresponding to "Deck A" in the original IGT is always shown in blue, "Deck B" in red, "Deck C" in purple, and "Deck D" in green. (E) RT distributions for the average subject (taking the mean RT at each trial in which a reach occurred to that

One of the primary advantages of the current task is that, in addition to manipulating the feedback schedules associated with target, the sensory uncertainty of the target itself could be manipulated. This was accomplished by changing the standard deviation of the target distribution, with higher standard deviations making it difficult to estimate the true target center (Körding & Wolpert, 2004). Given the normative predictions of the model, I predicted that increasing target estimation uncertainty would increase exploratory dynamics in the trial-wise reaching behavior (see Figure 7B). Figure 17A-B shows example reaches for two subjects, highlighting the reaching dynamics across 200 trials under high and low sensory uncertainty. The first subject experienced the low variance condition in the first block, shown as the left set of reaches with smaller target circles. In both the low and high variance conditions, this subject displayed a preference for targets A (blue), B (red), and D (green) and largely avoided target C (purple). While the Target preferences remained largely the same, the reach trajectories to each

preferred target became noticeably less precise in the high variance condition (right). The second subject, shown in Figure 17B, showed a strong preference for target D in both high (left) and low (right) uncertainty conditions, but reached to all four targets noticeably more when target variance was high (Figure 17B, left) compared to low (Figure 17B, right). Also, similar to the first subject, the reach trajectories in the high Variance condition are contracted in comparison to those in the low Variance condition. As a formal measure of target exploration, I calculated the probability of switching to a new reach target between trials (Figure 17C). This measure captures the assumption that exploration of different targets should yield a higher probability of reaching to a different target from trial to trial. In contrast with my original hypothesis, increasing target variance did not significantly impact the probability of switching; however, inspection of the reach trajectories did reveal a significant increase in the degree of endpoint error (Figure 17D) for high- compared to low-variance targets, $F(1, 32)= 30.11$, $p<.0001$. One interpretation is that the additional error in the high uncertainty condition is a reflection of poor estimation of the target center, rather than an indicator of active exploration, per se. Of course, it is also possible that this variability is capturing exploration of the reach space to facilitate convergence on an optimal trajectory.

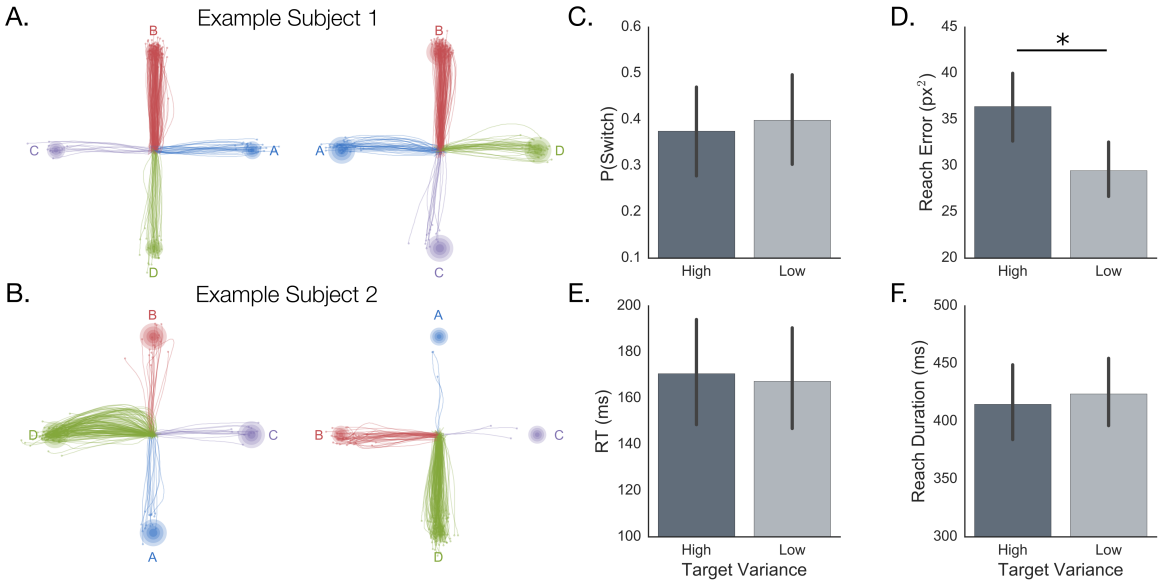

Figure 17. Target variance selectively increases reach error.

(A) Example Subject 1 showing reaching trajectories in the low variance (left) and high variance (right) conditions. The left-to-right order of the two sets of reaches depicts the order in which each subject experienced the high and low variance condition blocks. Reach traces are color-coded based on the feedback schedule assigned to each spatial location for that block of 200 trials. The Target with feedback schedule normally assigned to "Deck A" of the IGT is always shown in blue, "Deck B" in red, "Deck C" in purple, and "Deck D" in green. (B) Reach trajectories for a second example subject in the high variance (left) and low variance (right) conditions. Subject-averaged means and 95% confidence intervals for the (C) lag-1 probability of switching to a new reach target (D) reach error, (E) RT, and (F) reach duration. Dark and light bars show the estimates corresponding to the high and low variance conditions, respectively.

4.2.2 Adaptive Multi-Alternative Accumulator Model

Methods for optimizing parameters of accumulator models, such as those used to fit the static DPM in 3.2.3 have become widely adopted as the gold standard for testing model-based hypotheses. However, in order for parameter optimization to be useful, there are many assumptions that must first be validated about the statistical influence of the model's parameters, (covariance, effect-sizes, etc.). For the present purposes, I have chosen to forego a formal optimization of the network's parameters. Instead I will examine how simple predictions arise

from hypothesis-driven manipulation of parameters representing different forces on the activity of each action's direct and indirect pathway in a simplified decision network.

I first performed a grid search of the learning parameters to investigate the effect of Go (α^G) an NoGo (α^N) learning rates on the network's Payoff, and how this relationship was influenced by the inverse temperature parameter (β). Each plot in Figure 18B shows the networks *P* score for a single value of α^N and different combinations of α^G and β. In each plot, lighter lines show simulations with low values of α^G (lowest = .01) with darker lines reflecting increasingly high values of α^G (highest = .4). Moving from the left to right plots, *P* is shown to decrease as the ratio of a α^G to α^N increases, exacerbated by higher β values at each level. This relationship suggests that Payoff is modulated as a function of an agent's relative sensitivity to positive and negative feedback.

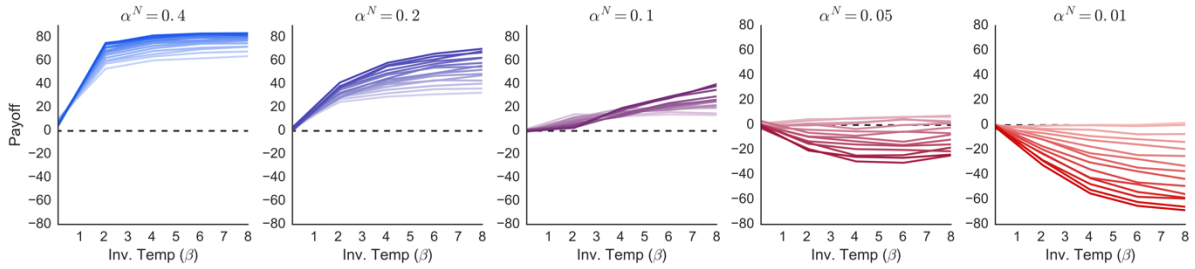

Figure 18. Payoff as a function of different learning strategies

Payoff scores for network IGT simulations. Each line reflects the average results from 100 simulated agents using a unique combination of Go (α^G) and NoGo (α^N) learning rates Go learning rates (α^G) ranging from 0.01 (lighter lines) to 0.4 (darker lines), and softmax temperature parameter (β). Each network simulation followed the same trial procedures as participants in the IGT described above (200 total trials with same feedback schedules as experimental subjects experienced). The following parameters were held constant across all simulations: boundary height (*a*= .4), non-decision time (*tr* = 100 ms), dynamic gain (γ =1.5), and initial drift-rates (eg. prior to learning) for the Go (ν^G=.7) and No Go (ν^N=.4) decision variables.

While learning from positive feedback is surely important, these simulations suggest that long term gains suffer to a greater extent when adaptation to negative feedback is weak. For instance, the rightmost plot shows how long-term outcomes suffer as a result of impaired learning from negative feedback, and how matters become worse as sensitivity to positive feedback increases (darker red lines). The rate of decay across values of β in each plot depict how "exploited" a given combination of α^G and α^N are. In other words, when actively exploiting a strategy of learning only from positive feedback, long-term outcomes become increasingly negative. Conversely, in the leftmost plot, exploiting a strategy of high positive and negative feedback has the exact opposite effect – maximizing long term gains with increasing levels of β.

In section 2.3 I speculated that RT should become faster as learning strategy becomes less exploratory and more exploitative (see Figure 7B). A general assumption within the Believer-Skeptic framework is that the dopaminergic reinforcement of direct pathway for high-value actions should not only promote that action but also increase the speed of execution. To test this assumption in the context of the IGT, I ran a simulation with 40 agents using the same feedback schedules as the actual experiment and asked how P score related to RTs, as a measure of a speed-accuracy tradeoff where, in this context, accuracy is determined as efficient value-based decision-making. The same correlation analysis was also performed for Q scores. Interestingly, I found that the simulated speed of responding was faster (r=.41, p=.02) for agents that strategically maximize long-term gains (e.g., higher P) but shows no linear relationship with the agent's sensitivity to reward frequency (r=.09, p>.05).

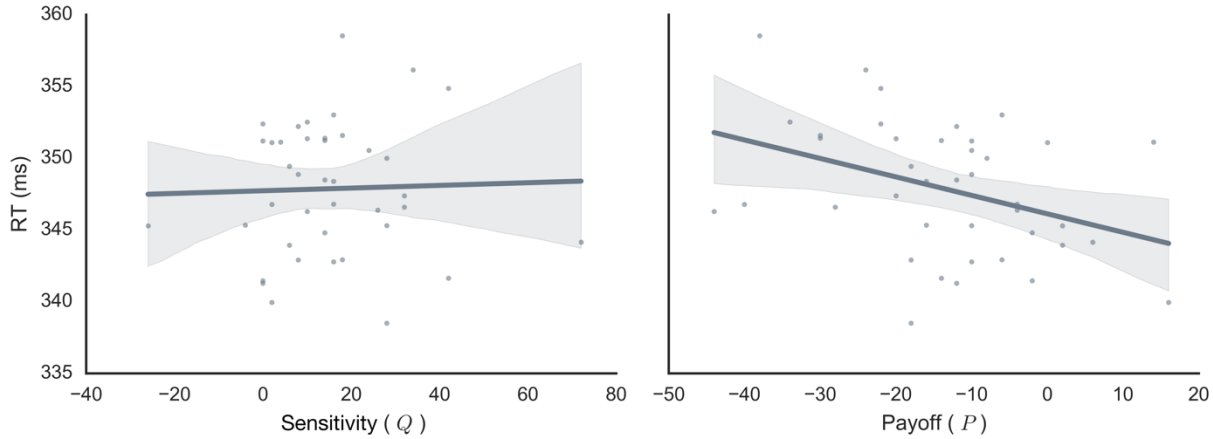

Figure 19. Simulated agents show negative correlation between RT and Payoff, not Sensitivity

Simulations of 40 agents with α^G to α^N both set equal to .2 and β=5. (medium blue line in second plot from the left in Figure 18). Correlations between Sensitivity (Q, left) and Payoff (P, right) with Response Time (RT) were calculated based on the simulated averages for each agent. Despite all having the same learning parameters, individual differences naturally arise due to the stochastic nature of both the model and the task. Thus, this simulation captures similar assumptions as when done with empirical data.

The model prediction in Figure 19 reveals a signature of the exploration-exploitation trade-off. Agents in a strongly exploitative state should have faster RTs (see Figure 7B), but being too exploitative biases you to immediate feedback signals and reduces the efficiency of long-term value based decisions. In contrast, agents that are more exploratory (i.e., longer RTs) tend to make more efficient long-term decisions. To test this model prediction against the observed data I calculated P and Q measures for each subject and correlated these scores with RT, defined as the time between cue offset and movement initiation, as well as the reach duration. Indeed, subjects with high P scores had significantly faster RTs (r=-.430, p<.05) and made faster reaches (r=-.547, p<.001) than lower scoring subjects (Figure 20, right), whereas correlating the same behavioral measures with subject-estimated Q scores revealed no reliable relationship (Figure 20, left). Thus, subjects who adopt the optimal strategy of maximizing long-term gains also appear to maximize their efficiency by gaining speed.

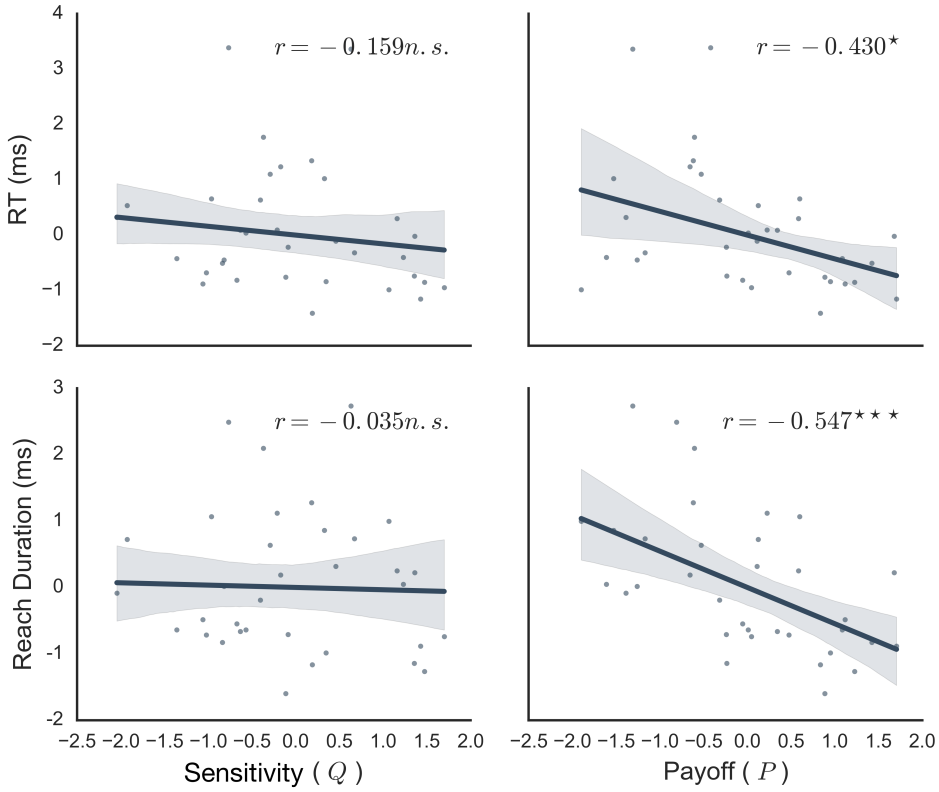

Figure 20. Payoff and Sensitivity have dissociable behavioral signatures.

Sensitivity (Q, left column) and Payoff (P, right column) were calculated for each subject and correlated with the subjects average RT (top row) and reach duration (bottom row). The results offer support for the model predictions in Figure 18, revealing a selective benefit on choice and movement efficiency for optimal strategies that exploit long term-gains.

Another empirical result of interest in the current study is the finding that many subjects preferred target B, which yields frequent (9+:1-) and high positive rewards (+100) and extreme, but infrequent penalties (-1250). This outcome has surfaced in a number of recent studies and has become central to the debate regarding the original predictions of the IGT for normal, healthy adults – that they exercise optimal decision-making by prioritizing actions that maximize long-term gains (C & D) over risky alternatives (A & B). Target A is easier to avoid as it has both a low frequency of reward as well as negative long-term payout. Conversely, target B provides long runs of high-value rewards before issuing any penalties. Several clinical studies have shown

that patients with Schizophrenia are more apt to exhibit prominent deck B selections, often attributed to distorted representations of value (Shurman et al., 2005). Other studies, however, where subjects are surveyed about explicit awareness of "good" and "bad" decks, have found that those who exhibit a preference for deck B report knowing the relative risks involved (Chiu & Lin, 2007). Thus, there are likely at least two different mechanisms underlying this strategy. To investigate the underlying mechanisms, I examined the combinations of α^G, α^N and β from the previous grid search that led to suboptimal target B selections and identified two parameter profiles that produced this effect.

Interestingly, both profiles led to similar patterns in RT across targets (Figure 21A-B, upper-left), in line with the current findings showing no significant mean different in RT across targets. Based on the selection count for each target (Figure 21A-B upper-right), however, the current results appear more similar to the first scenario (Figure 21B), in which B and D are preferred, reflecting a high sensitivity to reward frequency at the expense of long-term outcomes. This pattern arises when α^G and α^N are similarly matched but in the context of a low β weight, thus a more exploratory and less exploitative strategy. The second scenario matches studies showing clinical impairments in obesity (Brogan, Hevey, & Pignatti, 2010), Schizophrenia (Shurman et al., 2005), and ADHD (Toplak et al., 2005) in which a sub-optimal strategy of discounting negative feedback (low α^N) and over-weighting rewards (high α^G) is actively exploited (high β).

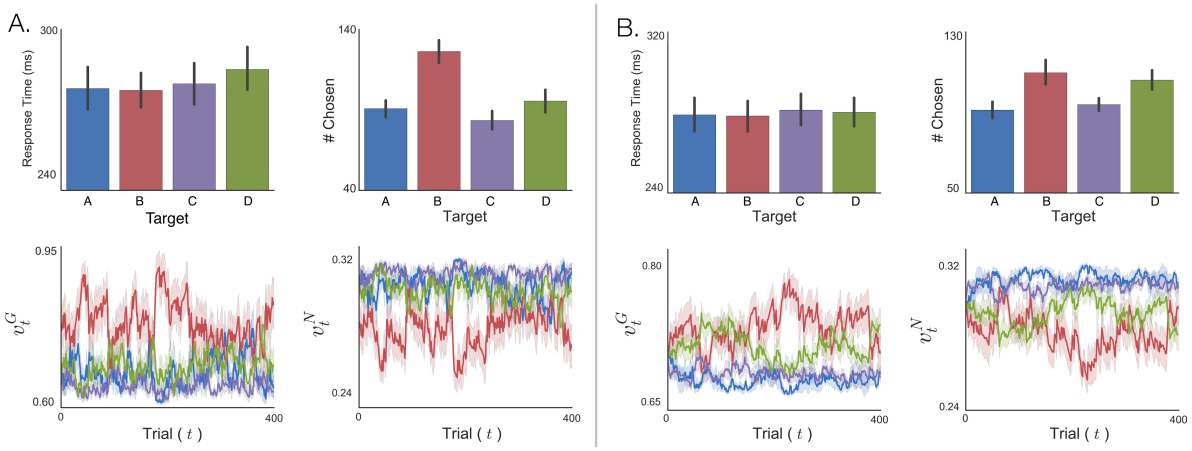

Figure 21. Simulations of prominent "Deck B" phenomenon

(**A**) Simulations with 40 agents with α^N=.01, α^G=.4, β=8 and (**B**) α^N=.2, α^G=.2, β=2. RT (upper left) and number of choices (upper right) from each target. The bottom row shows for each set of simulations the change in v^G and v^N across all simulated trials. In **A**, target B is exploited early on with an insensitivity in α^N leading to diminished effects of negative feedback on v^G and v^N.

4.3 SUMMARY OF RESULTS

In the current study, I developed a novel behavioral task for studying the interaction between reward-based learning and sensorimotor adaptation in a virtual reaching paradigm, where the feedback associated with each target was modeled after a deck from the IGT. The IGT is widely regarded as the keystone paradigm for studying complex, economic decision-making and has become a mainstay in both basic (Horstmann et al., 2012; Steingroever et al., 2013; Worthy et al., 2013) and clinical cognitive science (Fellows, 2007; Fellows & Farah, 2005; Sescousse, Barbalat, Domenech, & Dreher, 2013). The primary aim of the IGT is to isolate a decider's sensitivity to reward frequency from their preference for maximizing long-term gains. Typically, this involves tallying the number of draws from different combinations of the four decks, then using these counts to derive an estimate of preference for one feedback dimension or the other (Bechara et al., 1994). However, the collection of a single behavioral measure places a harsh

upper limit on making inferences about the cognitive and/or neural mechanisms that mediate economic choice.

To overcome this limitation, I combined the well-designed feedback schedules from the original IGT with a sensorimotor reaching task that allowed for rich behavioral assay of the temporal and kinematic effects of reinforcement, along dimensions of frequency and value. By investigating economic decisions in the context of a sensorimotor experiment, this opened up the door for manipulations of other types of uncertainty that could potentially compound those addressed in the original IGT. In the present study, I hypothesized that increasing the sensory uncertainty in the target location would force subjects into a more exploratory state and away from exploiting a given strategy. To encourage subjects to prioritize accurate target estimation, the original feedback score following each reach was scaled relative to how close the reach endpoint was to the center of the target. The behavioral results indicated that increasing sensory uncertainty did not have any effect on the decision of which target to reach to on each trial, as initially predicted, but did selectively increase exploration of the final reach location. Future studies should test the limits of this relationship as it is possible that the high variance condition in the present study did not sufficiently impair target estimation, leaving intact the ability to exploit the same strategies used by the subject in the low variance condition.

For all the reasons mentioned above, this task provided a unique opportunity for testing several key predictions laid out in Chapter I within the broader Believer-Skeptic framework. Here I focused on the prediction that exploitation of dopaminergic learning signals should promote advantageous decision-making by increasing the ratio of direct to indirect pathway activation in respective action channels of the BG, driving up the rate of evidence accumulation to facilitate faster action execution. To this end, I constructed a simple network model

representing each target as a single action channel driven by weighted input to functionally opposing Go and NoGo units that converged in the output to compete for action control. The network was repeatedly simulated using the same choice-dependent feedback schedules as in the experiment under parameter schemes that simulated varying reliance on exploratory and exploitative strategies as well as plausible differences in sensitivity to positive versus negative reinforcement. I then investigated how "individual differences" in Payoff and RT varied across simulated agents within a single parameter scheme, showing the predicted relationship between these dimensions was sustained in the current network setup and in the context of the IGT. Finally, I confirmed this prediction in the observed data by showing that individual differences in response speed and reach velocity were positively correlated with the exploitation of long-term gains. response s and reach duration. I discuss additional details and implications of this study in the next and final section.

5.0 FINAL SUMMARY AND CONCLUSIONS

Over the past decade, extraordinary progress has emerged from the joining of computational modeling with experimental approaches to cognitive neuroscience. This is especially true for the two computational frameworks addressed in this thesis: (1) reinforcement learning theory, a major catalyst in the discovery of dopamine-encoded prediction errors (Schultz, 2015) and numerous corollary advancements; and (2) accumulator models of decision making, that were revived by the seminal discovery of single neurons in parietal cortex showing the same stochastic rise-to-threshold dynamics during simple perceptual choice experiments (Gold & Shadlen, 2007). Both frameworks have solidified the importance of formal models in the study of cognitive phenomena, creating a quantitative link between mental processes and behavior, and more recently, brain activity.

In this dissertation, my aims were to synthesize RL theory with accumulation-to-bound models of decision making in ways that 1) satisfy certain assumptions about how these functions arise from cortico-BG circuitry and 2) address critical aspects of behavior that emerge from experience-driven changes in the dynamics of choice. In the remaining sections I revisit the key issues addressed in each chapter and discuss how they fit into the broader literature on BG involvement in adaptive decision-making, as well as limitations of each study. Finally, I discuss new avenues of work in this domain that I believe are worth pursuing.

5.1 ENCODING UNCERTAINTY AS A COMPETITION

In Chapter I, I proposed a novel framework for thinking about how the BG contribute to decision-making and identified points of convergence between the neural substrates involved in deciding and learning from environmental feedback. There I focused on three major points – first, highlighting newly discovered features of cortico-BG circuitry and their hypothesized role in generating adaptive behavior, specifically in the context of uncertainty. Drawing on evidence from recent electrophysiological work in animal models, behavioral and neuroimaging studies in humans, and computational theory, I proposed that the primary function of the BG is to guide action decisions in a flexible and goal-directed manner by maintaining control over decision certainty. This assertion relies on a fundamental assumption that the direct "Go" and indirect "NoGo" pathways act as competing forces on an action decision, with activation of the direct pathway playing the role of the Believer and that of the indirect pathway the role of the Skeptic (Dunovan et al., 2015).

Supporting evidence for the co-activation of the direct and indirect pathways was recently provided by Cui et al., (2013) showing that both pathways are engaged leading up the execution of an action. According to the canonical model of the BG, each action is represented as a channel, and each channel is composed of a direct and indirect pathway through which cortical commands may facilitate or suppress that action, respectively. In this model, the direct and indirect pathways act as independently operated levers: direct pathway to facilitate and indirect pathway to suppress. Recent computational work by Gurney et al., (2015), however, showed that independent operation of one pathway or the other leads to over-activation of too many channels at once, or a blanket suppression over all motor output. These authors found that the solution to this all or none problem was to allow cortical input to drive activation in both pathways for all

action channels being considered for execution, resulting in actions being driven by competition in the output of the BG where the two cortico-striatal pathways converge in the GPi. Using a simplified attractor network to simulate the competition between the direct and indirect pathways, I showed how this basic principle could act as a unifying thread for describing the functional roles of BG in action control, decision-making, and RL. Expanding on a model proposed by Dunovan et al., (2015), I showed that increasing activation of the Skeptic exerts proactive control over an action by overriding the Believer and suppressing the rate of evidence accumulation. Critically, this graded competition allows for varying degrees of uncertainty to be expressed by delaying execution to allow more cortical evidence to accumulate or by suppressing the action all together.

Given this mechanism for encoding uncertainty in a single action, I next examined how multiple actions could be represented in the context of a binary perceptual decision task, showing how changes in background excitability could modulate the speed-accuracy tradeoff. This simulation added support to a growing body of human neuroimaging studies showing that cortico-striatal activation increases to emphasize decision speed (Forstmann et al., 2008; Forstmann, Anwander, et al., 2010; van Maanen et al., 2016) or facilitate expected actions (Forstmann, Brown, et al., 2010). These studies have routinely shown that speed-related activation in the striatum is coupled with, and likely driven by control signals in the preSMA (Forstmann et al., 2008; King et al., 2012), which delivers diffuse excitatory drive to broad portions of the striatum. However, there remains no clear mechanistic hypothesis for how such unfocused excitation to both striatal pathways could selectively encourage speed over accuracy in the decision process. Counter to the intuition that such connectivity would be canceled out by exciting both direct and indirect populations, the model showed increasing the background

excitability had an accelerating effect on competing Believer-Skeptic competition. Future studies will be needed to validate this prediction of the model at the cellular and circuit levels. For instance, multi-site recordings from the striatum, GPe, and GPi would provide important insights into the action-channel hypothesis presumed by Believer-Skeptic. This is a core assumption of many BG-related models that rests on relatively weak empirical support. The theoretical framework I proposed in Chapter I offers both a new way of conceptualizing existing physiological studies of the BG pathways and novel predictions for future empirical work.

5.2 MECHANISMS UNDERLYING CONTEXTUAL CONTROL

A basic assumption of the Believer-Skeptic framework is that the BG acts as a bridge for binding cognitive processing (e.g., current goals, prior knowledge, perceptual judgements, etc.) with the appropriate action. In Chapter I, I covered two large bodies of literature on opposite sides of this bridge – on one side, studies of inhibitory control have identified critical opponent process dynamics between the direct and indirect pathways that give rise to efficient action selection and rapid cancellation of in response to unexpected shifts in environment; - and on the other side, studies of RL show how dopaminergic signals train cortico-striatal connections to associate valuable features in the environment with the actions that aid in achieving simple goals. Despite making up a generous portion of the behavioral research on the BG, inhibitory control and RL literatures have largely proceeded in parallel.

In section 3.0 , my goal was to test the prediction that, in addition to mediating value-based learning, dopaminergic reinforcement of the direct and indirect pathways should act as a tutor for proactively adjusting control to meet varying task demands. In a previous study, I

proposed a variation on the classic independent race model. While the IRM has been found to successfully predict behavior in a variety of stop-signal tasks, the notion of fully independent control signals underlying action execution and cancellation conflicts with the known overlap in their neural circuitry. In previous work (Dunovan et al., 2015) I showed that, by adapting the assumptions of the model to account for the convergence of proactive Go-NoGo and Braking signals in the BG output, the DPM was able to capture proactive and reactive forms of inhibition as well as task-evoked fMRI activation in premotor and thalamic regions. I found that the model was best able to capture behavioral and neural correlates of control by allowing contextual cues to proactively modulate the execution drift-rate, slowing responses to afford greater stopping efficacy by the nested Braking process. Following from this result, I developed an adaptive version of the DPM to test whether this same mechanism could explain gradual changes in behavior as the control demands of the environment are learned through experience.

The results of this experiment added supporting evidence for a contextual modulation of the drift-rate parameter. Furthermore, by modeling the effect of trialwise feedback about the temporal precision of responses (i.e., with respect to the Target RT) on Go trials as well the Context-dependent statistics of the stop-cue on Stop trials, this simple mechanism was able to account for the timecourse of learning multiple dimensions of behavior.

The findings presented in section 3.0 offer one potential solution to the issue of representing time in computational models of RL. Current RL models have largely focused on choice outcome. Focusing on a single dimension of behavior such as choice can be useful when testing model-predictions that expressly target that dimension, and other times this is not a choice but a necessary tradeoff with simplicity. However, when possible, testing predictions against RT data provides valuable constraints on the space of cognitive algorithms to consider

when comparing different models. Finally, with respect to behavioral models of the BG, recent optogenetic studies have begun probing cortico-striatal involvement in shaping movement kinematics, such as velocity and amplitude, inviting exciting new questions that push the boundaries current theories of BG-mediated behavioral control (Dudman & Krakauer, 2016; Wang et al., 2013; Yttri & Dudman, 2016). The model framework I proposed in section 3.0 offers a computational perspective by which to predict manipulations of dopaminergic pathways during learning.

5.3 ADAPTING MULTI-ALTERNATIVE DECISIONS TO FEEDBACK

The present state of computational modeling in human and animal decision-making remains overwhelmingly concerned with the dynamics of unitary and binary decisions, and is rooted mostly in assumptions of perception (Gold & Shadlen, 2007). This perseveration is not without cause, and abandoning perception as a model system for studying choice dynamics would be counterproductive to say the least. While there are still a host of unresolved questions surrounding the nature of perceptual signals in decision-making (e.g., how I decide if the dots are moving to the left or the right in the first place, let along how does this impact decision dynamics), I do believe it is a worthwhile exercise to examine how our current models stand up in bigger arenas of multi-alternative choices beyond two choices.

In section 4.0 , I set out to investigate how exploration-exploitation tradeoffs arise in the context of a multi-alternative reaching experiment, with target feedback designed to mirror that of different cards in the IGT. The exploration-exploitation tradeoff is a useful construct that characterizes two behavioral policies for navigating uncertainty, and often must be alternated

depending on context (Hills, Todd, Lazer, Redish, & Couzin, 2014). Take the example of learning to play the guitar. The first time a student picks up the instrument they automatically engage in exploration to find the most comfortable position to hold the neck, different ways to angle their wrist, how to rest their picking hand, and so on. Each of these positions undergoes a parameter optimization process, all converging toward the common goal of minimizing the initial awkwardness of holding a guitar. Over time, the student learns to exploit the sweet spots along each dimension until picking up the instrument becomes entirely automatic.

Both the behavioral and neural correlates of this tradeoff have been thoroughly studied in song-birds. In one particularly elegant study, Woolley et al., (2014) found that juvenile males displayed wide ranging variability in the temporal sequencing and frequencies of vocal production in isolation. When in the presence of a potential mate, however, songs became highly precise versions of the noisier rendition performed when alone. This observation was paired with neural recordings in the avian analog of the basal ganglia showing that the stochastic nature of firing pattern in the BG output nucleus observed during isolated performances became highly predictable when performing for a mate. Human neuroimaging studies have also found neural signatures of policy switching the human striatum (Li & Daw, 2011), and have shown that multivariate patterns in the caudate nucleus track uncertainty across state changes in volatile environments (Jiang, Beck, Heller, & Egner, 2015).

In section 2.0 , I proposed several mechanisms by which cortico-BG circuits could enforce more exploratory or exploitative learning policies, for instance increasing he level of tonic dopamine should dampen sensitivity to feedback by saturating receptor occupancy in the striatum, thereby preventing plasticity in the current cortico-striatal weighting. Similarly, exploitative behavior could be facilitated by increasing the background excitation through

diffuse glutamatergic inputs, such as those delivered by preSMA. Both of these mechanisms led to a common behavioral prediction – that the execution of exploited actions should be faster. In 4.0 , I confirmed that this prediction was upheld in a simple four-channel network model, each channel representing a target in the reaching task as a competing pair of Go and NoGo units, showing that variability in Payoff across simulated agents yielded a positive relationship with RT. Finally, I confirmed this prediction by showing how subjects that maximized Payoff by exploiting actions with long-term gains also exhibited speeded RTs and reach speed compared to less exploitative subjects in a multi-alternative reaching paradigm inspired by the IGT.

5.4 LIMITATIONS

The body of work presented in this dissertation has several limitations that warrant attention. In 4.0 , I discuss a wide range of possible, but largely speculative, links between the circuit-level properties of cortico-BG networks that give rise to complex behavioral phenomena. Further to the point, the Believer-Skeptic simulations presented in 4.0 , that form the crux of many corollary arguments made throughout the document, offers a comprised definition of biological plausibility, trading realistic complexity for behavioral understandability. The elegance of the algorithmic models proposed in this work is that they are constrained by biological assumptions, but the nature of these constraints in the decision process can only be confirmed through circuit-level analysis of the BG pathways during adaptive decision-making. So, despite being premised off of a large body of emerging physiological studies, the specific biological extensions of this work remain relatively untested.

In section 3.0 , the between-group design precluded fits of the static model to individual subject data. This does not invalidate the results by any means, as group-level fits have been found to be reliable approximations of aggregated fits performed at the subject-level. Moreover, the Adaptive model was trained on individual data sets so as to reliably capture the rate of drift-rate and T^S adaptation across the span of a single experimental session. Still, future studies aimed at addressing similar learning hypotheses in accumulator dynamics would benefit from a repeated measures design, pending the absence of confounding order effects across conditions.

Finally, in comparison with the more rigorous dissection of the behavior in section 3.0 , the link between the network model and reaching dynamics in section 4.0 offers a more tenuous link between a more complex model and a richer dataset. The model presented was simply used as a predictor for normative behavioral patterns. The increased model complexity quickly expands the number of free variables to formally fit to individual data. Further work can and should be done to identify efficient model fitting methods for dealing with more highly-parameterized models like that described in section 4.1.3.

5.5 CONCLUSION

Computational models of decision making and reinforcement learning have been veritable workhorses in the fields of cognitive psychology and neuroscience, providing formal quantitative descriptions of two of the most fundamental aspects of behavior. Here I have argued for a biologically motivated synthesis of these two frameworks, motivated by their shared implementation in cortico-BG circuitry. This approach adheres to David Marr's (1982) classic proposal that a true understanding of intelligent behavior requires an integrative analysis of its

implementational (e.g., neurobiological) and algorithmic (e.g., cognitive) levels. In spite of the limitations discussed above, I believe that this work has taken a meaningful step forward toward understanding how the dynamics of decision-making change with experience as well as the neural systems that mediate this phenomenon.

BIBLIOGRAPHY

Abdi, H. (2010). The greenhouse-geisser correction. *Encyclopedia of Research Design. Sage Publications*, 1–10. doi:10.1007/BF02289823

Albin, R. L., Young, A. B., & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, *12*(10), 366–75.

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*, 357–381. doi:10.1146/annurev.neuro.9.1.357

Antoniades, C. a, Bogacz, R., Kennard, C., FitzGerald, J. J., Aziz, T., & Green, A. L. (2014). Deep brain stimulation abolishes slowing of reactions to unlikely stimuli. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(33), 10844–52. doi:10.1523/JNEUROSCI.1065-14.2014

Apicella, P., Ljungberg, T., Scarnati, E., & Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. *Experimental Brain Research*, *85*(3), 491–500. doi:10.1007/BF00231732

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to Stop signal response inhibition: role of the subthalamic nucleus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(9), 2424–33. doi:10.1523/JNEUROSCI.4682-05.2006

Aron, A. R., Robbins, T. W., & Poldrack, R. a. (2014). Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences*, *18*(4), 177–85. doi:10.1016/j.tics.2013.12.003

Bahuguna, J., Aertsen, A., & Kumar, A. (2015). Existence and Control of Go/No-Go Decision Transition Threshold in the Striatum. *PLOS Computational Biology*, *11*(4), e1004233. doi:10.1371/journal.pcbi.1004233

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, *27*(31), 8161–5. doi:10.1523/JNEUROSCI.1554-07.2007

Barbera, G., Liang, B., Zhang, L., Gerfen, C. R., Culurciello, E., Chen, R., … Lin, D.-T.

(2016). Spatially Compact Neural Clusters in the Dorsal Striatum Encode Locomotion Relevant Information. *Neuron*, 202–213. doi:10.1016/j.neuron.2016.08.037

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15. doi:10.1016/0010-0277(94)90018-3

Bertucco, M., Bhanpuri, N. H., & Sanger, T. D. (2015). Perceived cost and intrinsic motor variability modulate the speed-accuracy trade-off. *PLoS ONE*, *10*(10), 1–18. doi:10.1371/journal.pone.0139988

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), 700–65. doi:10.1037/0033-295X.113.4.700

Bogacz, R., & Cohen, J. D. (2004). Parameterization of connectionist models. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *36*(4), 732–741. doi:10.3758/BF03206554

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*(2), 442–77. doi:10.1162/neco.2007.19.2.442

Bogacz, R., & Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural Computation*, *23*(4), 817–851. doi:10.1162/NECO_a_00103

Bogacz, R., Wagenmakers, E., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–6. doi:10.1016/j.tins.2009.09.002

Bolam, J. P., Hanley, J. J., Booth, P. a, & Bevan, M. D. (2000). Synaptic organisation of the basal ganglia. *Journal of Anatomy*, *196 (Pt 4*(September 2000), 527–542. doi:10.1046/j.1469-7580.2000.19640527.x

Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. *Nature*, *417*(May), 351–358.

Brogan, A., Hevey, D., & Pignatti, R. (2010). Anorexia, bulimia, and obesity: shared decision making deficits on the Iowa Gambling Task (IGT). *Journal of the International Neuropsychological Society*, *16*(4), 711–5. doi:10.1017/S1355617710000354

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–78. doi:10.1016/j.cogpsych.2007.12.002

Cazorla, M., de Carvalho, F. D., Chohan, M. O., Shegda, M., Chuhma, N., Rayport, S., …

Kellendonk, C. (2014). Dopamine D2 receptors regulate the anatomical and functional balance of basal ganglia circuitry. *Neuron*, *81*(1), 153–64. doi:10.1016/j.neuron.2013.10.041

Chiu, Y.-C., & Lin, C.-H. (2007). Is deck C an advantageous deck in the Iowa Gambling Task? *Behavioral and Brain Functions : BBF*, *3*(37), 11. doi:10.1186/1744-9081-3-16

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*(6), 693–702. doi:10.1038/nn0708-851c

Cockburn, J., Collins, A. G. E., & Frank, M. J. (2014). A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. *Neuron*, *83*(3), 551–557. doi:10.1016/j.neuron.2014.06.035

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), 337–66. doi:10.1037/a0037015

Cox, S. M. L., Frank, M. J., Larcher, K., Fellows, L. K., Clark, C. A., Leyton, M., & Dagher, A. (2015). Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. *NeuroImage*, *109*, 95–101. doi:10.1016/j.neuroimage.2014.12.070

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), 238–42. doi:10.1038/nature11846

Dean, M., Wu, S.-W., & Maloney, L. T. (2007). Trading off speed and accuracy in rapid, goal-directed movements. *Journal of Vision*, *7*(5), 10.1-12. doi:10.1167/7.5.10

Ding, L., & Perkel, D. J. (2014). Two tales of how expectation of reward modulates behavior. *Current Opinion in Neurobiology*, *29*, 142–147. doi:10.1016/j.conb.2014.07.011

Drugowitsch, J., Deangelis, G. C., Angelaki, D. E., & Pouget, A. (2015). Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *Elife*, 1–11. doi:10.7554/eLife.06678

Dudman, J. T., & Krakauer, J. W. (2016). The basal ganglia: From motor commands to the control of vigor. *Current Opinion in Neurobiology*, *37*, 158–166. doi:10.1016/j.conb.2016.02.005

Dunovan, K., Lynch, B., Molesworth, T., & Verstynen, T. (2015). Competing basal-ganglia pathways determine the difference between stopping and deciding not to go. *eLife*, 1–24. doi:10.7554/eLife.08723

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*(March), 1–15. doi:http://dx.doi.org/10.1101/037085

Fellows, L. K. (2007). The role of orbitofrontal cortex in decision making: a component process account. *Annals of the New York Academy of Sciences*, *1121*, 421–30. doi:10.1196/annals.1401.023

Fellows, L. K., & Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex (New York, N.Y. : 1991)*, *15*(1), 58–63. doi:10.1093/cercor/bhh108

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., … Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(36), 15916–20. doi:10.1073/pnas.1004932107

Forstmann, B. U., Brown, S., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: a model-based analysis. *Frontiers in Human Neuroscience*, *4*(May), 40. doi:10.3389/fnhum.2010.00040

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17538–17542. doi:10.1073/pnas.0805903105

Forstmann, B. U., Keuken, M. C., Jahfari, S., Bazin, P.-L., Neumann, J., Schäfer, A., … Turner, R. (2012). Cortico-subthalamic white matter tract strength predicts interindividual efficacy in stopping a motor response. *NeuroImage*, *60*(1), 370–5. doi:10.1016/j.neuroimage.2011.12.044

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*(1), 51–72. doi:10.1162/0898929052880093

Frank, M. J., Doll, B. B., Oas-terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. doi:10.1038/nn.2342

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V, Cavanagh, X. F., & Badre, D. (2015). fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *The Journal of Neuroscience*, *35*(2), 485–494. doi:10.1523/JNEUROSCI.2036-14.2015

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.)*, *306*(5703), 1940–3. doi:10.1126/science.1102941

Friedman, A., Homma, D., Gibb, L. G. G., Amemori, K., Rubin, S. J. J., Hood, A. S. S., … Graybiel, A. M. M. (2015). A Corticostriatal Path Targeting Striosomes Controls

Decision-Making under Conflict. *Cell*, *161*(6), 1320–1333. doi:10.1016/j.cell.2015.04.049

Friend, D. M., & Kravitz, A. V. (2014). Working together: basal ganglia pathways in action selection. *Trends in Neurosciences*, *37*(6), 301–3. doi:10.1016/j.tins.2014.04.004

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(30), 535–74. doi:10.1146/annurev.neuro.29.051605.113038

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. doi:10.1371/journal.pbio.1002034

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience*, *34*(3), 698–704. doi:10.1523/JNEUROSCI.2489-13.2014

Hawthorne, M. J., & Pierce, B. H. (2015). Disadvantageous Deck Selection in the Iowa Gambling Task: The Effect of Cognitive Load. *Europe's Journal of Psychology*, *11*(2), 335–348. doi:10.5964/ejop.v11i2.931

Herz, D. M., Zavala, B. A., Bogacz, R., & Brown, P. (2016). Neural Correlates of Decision Thresholds in the Human Subthalamic Nucleus. *Current Biology*, 1–5. doi:10.1016/j.cub.2016.01.051

Hikida, T., Yawata, S., Yamaguchi, T., Danjo, T., Sasaoka, T., Wang, Y., & Nakanishi, S. (2013). Pathway-specific modulation of nucleus accumbens in reward and aversive behavior via selective transmitter receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 342–7. doi:10.1073/pnas.1220358110

Hills, T. T., Todd, P. M., Lazer, D., Redish, a. D., & Couzin, I. D. (2014). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1). doi:10.1016/j.tics.2014.10.004

Hollerman, J. R., Tremblay, L., & Schultz, W. (1998). Influence of reward expectation on behavior-related neuronal activity in primate striatum. *J Neurophysiol*, *80*(2), 947–963. doi:10.1016/S0531-5131(03)00188-2

Horstmann, A., Villringer, A., & Neumann, J. (2012). Iowa Gambling Task: there is more to consider than long-term outcome. Using a linear equation model to disentangle the impact of outcome and frequency of gains and losses. *Frontiers in Neuroscience*, *6*(May), 1–10. doi:10.3389/fnins.2012.00061

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*(February), 1–14. doi:10.3389/fnins.2012.00009

Jahfari, S., Stinear, C. M., Claffey, M., Verbruggen, F., & Aron, A. R. (2010). Responding

with restraint: what are the neurocognitive mechanisms? *Journal of Cognitive Neuroscience*, *22*(7), 1479–92. doi:10.1162/jocn.2009.21307

Jahfari, S., Verbruggen, F., Frank, M. J., Waldorp, L. J., Colzato, L., Ridderinkhof, K. R., & Forstmann, B. U. (2012). How preparation changes the need for top-down control of the basal ganglia when inhibiting premature actions. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(32), 10870–8. doi:10.1523/JNEUROSCI.0902-12.2012

Jahfari, S., Waldorp, L., van den Wildenberg, W. P. M., Scholte, H. S., Ridderinkhof, K. R., & Forstmann, B. U. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(18), 6891–9. doi:10.1523/JNEUROSCI.5253-10.2011

Jiang, J., Beck, J., Heller, K., & Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nature Communications*, *6*(May), 1–11. doi:10.1038/ncomms9165

Kao, M. H., Doupe, A. J., & Brainard, M. S. (2005). Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. *Nature*, *433*(7026), 638–43. doi:10.1038/nature03127

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, Locus of Control, and the Exploration-Exploitation Tradeoff. *Neuropsychopharmacology*, *40*(2), 454–462. doi:10.1038/npp.2014.193

Keeler, J. F., Pretsell, D. O., & Robbins, T. W. (2014). Functional implications of dopamine D1 vs D2 receptors: A "Prepare and Select" model of the striatal direct vs. indirect pathways. *Neuroscience*, *282*(July), 156–175. doi:10.1016/j.neuroscience.2014.07.021

Keuken, M. C., Langner, R., Eickhoff, S. B., Forstmann, B. U., & Neumann, J. (2014). Brain networks of perceptual decision-making: an fMRI ALE meta-analysis. *Frontiers in Human Neuroscience*, *8*. doi:10.3389/fnhum.2014.00445

Keuken, M. C., Van Maanen, L., Bogacz, R., Schäfer, A., Neumann, J., Turner, R., & Forstmann, B. U. (2015). The subthalamic nucleus during decision-making with multiple alternatives. *Human Brain Mapping*, *0*(September). doi:10.1002/hbm.22896

King, A. V, Linke, J., Gass, A., Hennerici, M. G., Tost, H., Poupon, C., & Wessa, M. (2012). Microstructure of a three-way anatomical network predicts individual differences in response inhibition: a tractography study. *NeuroImage*, *59*(2), 1949–59. doi:10.1016/j.neuroimage.2011.09.008

Klanker, M., Feenstra, M., & Denys, D. (2013). Dopaminergic control of cognitive

flexibility in humans and animals. *Frontiers in Neuroscience*, *7*(7 NOV), 1–24. doi:10.3389/fnins.2013.00201

Körding, K. P., & Wolpert, D. M. (2004). Bayesian Integration in Sensorimotor Learning. *Nature*, *427*(5), 244–247. doi:10.1152/jn.00275.2004

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*(1), 287–308. doi:10.1146/annurev-neuro-062111-150512

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(14), 5504–11. doi:10.1523/JNEUROSCI.6316-10.2011

Lin, C.-H., Chiu, Y.-C., Lee, P.-L., & Hsieh, J.-C. (2007). Is deck B a disadvantageous deck in the Iowa Gambling Task? *Behavioral and Brain Functions : BBF*, *3*(16), 10. doi:10.1186/1744-9081-3-16

Lo, C.-C., Wang, C.-T., & Wang, X.-J. (2015). Speed-accuracy tradeoff by a control signal with balanced excitation and inhibition. *Journal of Neurophysiology*, *114*, jn.00845.2013. doi:10.1152/jn.00845.2013

Lo, C.-C., & Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, *9*(7), 956–63. doi:10.1038/nn1722

Majid, D. S. A., Cai, W., Corey-Bloom, J., & Aron, A. R. (2013). Proactive selective response suppression is implemented via the basal ganglia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(33), 13259–69. doi:10.1523/JNEUROSCI.5651-12.2013

Mallet, N., Micklem, B. R., Henny, P., Brown, M. T., Williams, C., Bolam, J. P., … Magill, P. J. (2012). Dichotomous Organization of the External Globus Pallidus. *Neuron*, *74*(6), 1075–1086. doi:10.1016/j.neuron.2012.04.027

Mallet, N., Schmidt, R., Leventhal, D., Chen, F., Amer, N., Boraud, T., … Boraud, T. (2016). Arkypallidal Cells Send a Stop Signal to Striatum Report Arkypallidal Cells Send a Stop Signal to Striatum. *Neuron*, 1–9. doi:10.1016/j.neuron.2015.12.017

Marcott, P. F., Mamaligas, A. A., & Ford, C. P. (2014). Phasic Dopamine Release Drives Rapid Activation of Striatal D2-Receptors. *Neuron*, *84*(1), 164–176. doi:10.1016/j.neuron.2014.08.058

Maritz, J. S., & Jarrett, R. G. (1978). the Variance of the A Note on Estimating Sample Median, *73*(361), 194–196. doi:10.1080/01621459.1978.10480027

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. *Vision: A computational investigation into the human*

representation and processing of visual information.

Marshall, J. A. R., Bogacz, R., Dornhaus, A., Planqué, R., Kovacs, T., & Franks, N. R. (2009). On optimal decision-making in brains and social insect colonies. *Journal of the Royal Society, Interface / the Royal Society*, *6*(40), 1065–74. doi:10.1098/rsif.2008.0511

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*(4), 381–425. doi:10.1016/S0301-0082(96)00042-1

Morita, K., & Kato, A. (2014). Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. *Frontiers in Neural Circuits*, *8*(April), 36. doi:10.3389/fncir.2014.00036

Mulder, M. J., van Maanen, L., & Forstmann, B. U. (2014). Perceptual decision neurosciences - A model-based review. *Neuroscience*, *277*(August), 872–884. doi:10.1016/j.neuroscience.2014.07.031

Murakami, M., Vicente, M. I., Costa, G. M., & Mainen, Z. F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. *Nature Neuroscience*, *17*(11), 1574–82. doi:10.1038/nn.3826

Nelder, B. J. a, & Mead, R. (1964). A simplex method for function minimization.

Oldenburg, I. A., & Sabatini, B. L. (2015). Antagonistic but Not Symmetric Regulation of Primary Motor Cortex by Basal Ganglia Direct and Indirect Pathways. *Neuron*, *86*(5), 1174–81. doi:10.1016/j.neuron.2015.05.008

Palmeri, T. J., Schall, J. D., & Logan, G. D. (2013). Neurocognitive Modeling of Perceptual Decision Making. *Oxford Handbook of Computational and Mathematical Psychology*. doi:10.1093/oxfordhb/9780199957996.013.15

Panigrahi, B., Martin, K. A., Li, Y., Graves, A. R., Vollmer, A., Olson, L., … Dudman, J. T. (2015). Dopamine Is Required for the Neural Representation and Control of Movement Vigor. *Cell*, *162*(6), 1418–1430. doi:10.1016/j.cell.2015.08.014

Parent, A., & Hazrati, L. (1995). Functional Anatomy of the basal ganglia. II. The place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Research Reviews*, *20*, 128–154.

Park, I. M., Meister, M. L. R., Huk, A. C., & Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, *17*(10), 1395–1403. doi:10.1038/nn.3800

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron*, *82*(3), 709–20. doi:10.1016/j.neuron.2014.03.014

Ratcliff, R. (1978). A theory of Memory Retrival. *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. doi:10.1162/NECO_a_00270

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–67. doi:10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *xx*(4), 1–22. doi:10.1016/j.tics.2016.01.007

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–81.

Schroll, H., & Hamker, F. H. (2013). Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy. *Frontiers in Systems Neuroscience*, *7*(December), 122. doi:10.3389/fnsys.2013.00122

Schultz, W. (2015). Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews*, *95*(3), 853–951. doi:10.1152/physrev.00023.2014

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32. doi:10.1038/nrn.2015.26

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(June 1994).

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500. doi:10.1146/annurev.neuro.23.1.473

Sescousse, G., Barbalat, G., Domenech, P., & Dreher, J. C. (2013). Imbalance in the sensitivity to different types of rewards in pathological gambling. *Brain*, *136*(8), 2527–2538. doi:10.1093/brain/awt126

Shadlen, M. N. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939. doi:10.1016/j.neuron.2016.04.036

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(2), 628–33.

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. *J Neurosci*, *34*(28), 9196–9201. doi:10.1523/JNEUROSCI.0313-14.2014

Shurman, B., Horan, W. P., & Nuechterlein, K. H. (2005). Schizophrenia patients demonstrate a distinctive pattern of decision-making impairment on the Iowa Gambling Task. *Schizophrenia Research*, *72*(2–3), 215–224. doi:10.1016/j.schres.2004.03.020

Silberberg, G., & Bolam, J. P. (2015). Local and afferent synaptic pathways in the striatal microcircuitry. *Current Opinion in Neurobiology*, *33*, 182–187. doi:10.1016/j.conb.2015.05.002

Simen, P. (2012). Evidence Accumulator or Decision Threshold - Which Cortical Mechanism are We Observing? *Frontiers in Psychology*, *3*(June), 183. doi:10.3389/fpsyg.2012.00183

Smith, Y., Bevan, M. D., Shink, E., & Bolam, J. P. (1998a). Microcircuitry of the Direct and Indirect Pathways of the Basal Ganglia. *Neuroscience*, *86*(2), 353–387. doi:http://dx.doi.org/10.1016/S0306-4522(98)00004-9

Smith, Y., Bevan, M., Shink, E., & Bolam, J. (1998b). Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience*, *86*(2), 353–387.

Standage, D., Blohm, G., & Dorris, M. C. (2014). On the neural implementation of the speed-accuracy trade-off. *Frontiers in Neuroscience*, *8*(August), 236. doi:10.3389/fnins.2014.00236

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). A Comparison of Reinforcement Learning Models for the Iowa Gambling Task Using Parameter Space Partitioning. *Journal of Problem Solving*, *5*(2), 1–32. doi:10.7771/1932-6246.1150

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, *30*(5), 228–235. doi:10.1016/j.tins.2007.03.008

Sutton, R. S., Barto, A. G., & Book, a B. (1998). Reinforcement Learning : An Introduction.

Tanaka, M. (2007). Cognitive Signals in the Primate Motor Thalamus Predict Saccade Timing. *Journal of Neuroscience*, *27*(44), 12109–12118. doi:10.1523/JNEUROSCI.1873-07.2007

Tanaka, M., & Kunimatsu, J. (2011). Contribution of the central thalamus to the generation of volitional saccades. *European Journal of Neuroscience*, *33*(11), 2046–2057. doi:10.1111/j.1460-9568.2011.07699.x

Taverna, S., Ilijic, E., & Surmeier, D. J. (2008). Recurrent Collateral Connections of Striatal Medium Spiny Neurons Are Disrupted in Models of Parkinson's Disease. *Journal of*

Neuroscience, *28*(21), 5504–5512. doi:10.1523/JNEUROSCI.5493-07.2008

Thura, D. (2016). How to discriminate conclusively among different models of decision making? *Journal of Neurophysiology*, *115*(5), 2251–2254. doi:10.1152/jn.00911.2015

Toplak, M. E., Jain, U., & Tannock, R. (2005). Executive and motivational processes in adolescents with Attention-Deficit-Hyperactivity Disorder (ADHD). *Behavioral and Brain Functions : BBF*, *1*(1), 8. doi:10.1186/1744-9081-1-8

Tremblay, L., Hollerman, J. R., & Schultz, W. (1998). Modifications of reward expectation-related neuronal activity during learning in primate striatum. *Journal of Neurophysiology*, *80*, 964–977.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*(2), 179–205. doi:10.1037/a0036137

Tumer, E. C., & Brainard, M. S. (2007). Performance variability enables adaptive plasticity of "crystallized" adult birdsong. *Nature*, *450*(7173), 1240–4. doi:10.1038/nature06390

Turner, R. S., & Desmurget, M. (2010). Basal ganglia contributions to motor control: a vigorous tutor. *Current Opinion in Neurobiology*, *20*(6), 704–16. doi:10.1016/j.conb.2010.08.022

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(48), 17488–95. doi:10.1523/JNEUROSCI.2924-11.2011

van Maanen, L., Fontanesi, L., Hawkins, G. E., & Forstmann, B. U. (2016). Striatal activation reflects urgency in perceptual decision making. *NeuroImage*, *139*(June), 294–303. doi:10.1016/j.neuroimage.2016.06.045

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22.

Wales, D., & Doye, J. P. K. (1997). Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *Journal of Physical Chemistry A*, *101*(97), 5111–5116. doi:10.1021/jp970984n

Wall, N. R., De La Parra, M., Callaway, E. M., & Kreitzer, A. C. (2013). Differential innervation of direct- and indirect-pathway striatal projection neurons. *Neuron*, *79*(2), 347–60. doi:10.1016/j.neuron.2013.05.014

Wang, A. Y., Miura, K., & Uchida, N. (2013). The dorsomedial striatum encodes net expected return, critical for energizing performance vigor. *Nature Neuroscience*, *16*(5),

639–47. doi:10.1038/nn.3377

Wiecki, T. V, & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*(2), 329–55. doi:10.1037/a0031542

Woolley, S. C., Rajan, R., Joshua, M., & Doupe, A. J. (2014). Emergence of context-dependent variability across a basal ganglia network. *Neuron*, *82*(1), 208–23. doi:10.1016/j.neuron.2014.01.039

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: a comparison of win-stay/lose-shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*(2), 364–71. doi:10.3758/s13423-012-0324-9

Yawata, S., Yamaguchi, T., Danjo, T., Hikida, T., & Nakanishi, S. (2012). Pathway-specific control of reward learning and its flexibility via selective dopamine receptors in the nucleus accumbens. *Proceedings of the National Academy of Sciences*, *109*(31), 12764–12769. doi:10.1073/pnas.1210797109

Yttri, E. A., & Dudman, J. T. (2016). Opponent and bidirectional control of movement velocity in the basal ganglia. *Nature*, *533*(7603), 1–16. doi:10.1038/nature17639