Behavioral/Systems/Cognitive

A Role for Dopamine in Temporal Decision Making and Reward Maximization in Parkinsonism

Ahmed A. Moustafa,[1]* Michael X. Cohen,[1] Scott J. Sherman,[2] and Michael J. Frank[1]*

[1]Department of Psychology and Program in Neuroscience, and [2]Department of Neurology, University of Arizona, Tucson, Arizona 85721

Converging evidence implicates striatal dopamine (DA) in reinforcement learning, such that DA increases enhance "Go learning" to pursue actions with rewarding outcomes, whereas DA decreases enhance "NoGo learning" to avoid non-rewarding actions. Here we test whether these effects apply to the response time domain. We employ a novel paradigm which requires the adjustment of response times to a single response. Reward probability varies as a function of response time, whereas reward magnitude changes in the opposite direction. In the control condition, these factors exactly cancel, such that the expected value across time is constant (CEV). In two other conditions, expected value increases (IEV) or decreases (DEV), such that reward maximization requires either speeding up (Go learning) or slowing down (NoGo learning) relative to the CEV condition. We tested patients with Parkinson's disease (depleted striatal DA levels) on and off dopaminergic medication, compared with age-matched controls. While medicated, patients were better at speeding up in the DEV relative to CEV conditions. Conversely, nonmedicated patients were better at slowing down to maximize reward in the IEV condition. These effects of DA manipulation on cumulative Go/NoGo response time adaptation were captured with our a priori computational model of the basal ganglia, previously applied only to forced-choice tasks. There were also robust trial-to-trial changes in response time, but these single trial adaptations were not affected by disease or medication and are posited to rely on extrastriatal, possibly prefrontal, structures.

Key words: **Parkinson's disease; basal ganglia; dopamine; reinforcement learning; computational model; reward**

Introduction

Parkinson's disease (PD) is a neurodegenerative disorder primarily associated with dopaminergic cell death and concomitant reductions in striatal dopamine (DA) levels (Kish et al., 1988; Brück et al., 2006). The disease leads to various motor and cognitive deficits including learning, decision making, and working memory, likely because of dysfunctional circuit-level functioning between the basal ganglia and frontal cortex (Alexander et al., 1986; Knowlton et al., 1996; Frank, 2005; Cools, 2006). Further, although DA medications sometimes improve cognitive function, they can actually induce other cognitive impairments that are distinct from those associated with PD itself (Cools et al., 2001, 2006; Frank et al., 2004, 2007b; Shohamy et al., 2004; Moustafa et al., 2008). Many of these contrasting medication effects have been observed in reinforcement learning tasks in which participants select among multiple responses to maximize their probability of correct feedback. Here we study the complementary role of basal ganglia dopamine on learning when to respond to maximize reward using a novel temporal decision making task. Although the "which" and "when" aspects of response learning might seem conceptually different, simulation studies show that the same neural mechanisms within the basal ganglia can support both selection of the most rewarding response out of multiple options, and how fast a given rewarding response is selected. This work builds on existing frameworks linking similar corticostriatal mechanisms underlying interval timing with those of action selection and working memory updating (Lustig et al., 2005), and further explores the role of reinforcement.

Various computational models suggest that circuits linking basal ganglia with frontal cortex support action selection (Berns and Sejnowski, 1995; Suri and Schultz, 1998; Frank et al., 2001; Gurney et al., 2001; Frank, 2006; Houk et al., 2007; Moustafa and Maida, 2007) and that striatal DA modulates reward-based learning and performance (Suri and Schultz, 1998; Delgado et al., 2000, 2005; Doya, 2000; Frank, 2005; Shohamy et al., 2006; Niv et al., 2007). In the models, phasic DA signals modify synaptic plasticity in the corticostriatal pathway (Wickens et al., 1996; Reynolds et al., 2001). Further, phasic DA bursts boost learning in "Go" neurons to reinforce adaptive choices, whereas reduced DA levels during negative outcomes support learning in "NoGo" neurons to avoid maladaptive responses (Frank, 2005) (see Fig. 2). This model has been applied to understand patterns of learning in PD patients (Frank, 2005), who have depleted striatal DA levels as a result of the disease, but increased striatal DA levels after DA medication (Tedroff et al., 1996; Pavese et al., 2006).

Supporting the models, experiments revealed that PD patients on medication learned better from positive than from negative reinforcement feedback, whereas patients off medication showed

Received July 4, 2008; revised Sept. 24, 2008; accepted Oct. 7, 2008.

This work was supported by National Institute on Mental Health Grant R01 MH080066-01. We thank Omar Meziab for help in administering cognitive tasks to participants, and two anonymous reviewers for helpful comments that led us to a deeper understanding of our data.

*A.A.M. and M.J.F. contributed equally to this work.

Correspondence should be addressed to Michael J. Frank at the above address. E-mail: mfrank@u.arizona.edu.

DOI:10.1523/JNEUROSCI.3116-08.2008

Table 1. Demographic variables for seniors and PD patients

Group	*n*	Sex ratio (m:f)	Age	Years of education	NAART (no. correct)	Hoehn and Yahr stage	Years diag
Seniors	17	7:10	65.6 (2.1)	16.5 (0.8)	39.8 (4.2)	N/A	N/A
PD patients	20	14:6	69.8 (1.5)	18.1 (0.8)	45.1 (1.8)	2.5 (0.5)	6.2 (1.2)

Groups were not gender-matched, but it is unlikely that this factor impacts on the results given that medication manipulations were within-subject. NAART, Number of correct responses (of 61) in the North American Adult Reading Test, an estimate of premorbid verbal IQ. For PD patients, disease severity is indicated in terms of mean Hoehn and Yahr stage, and the number of years since having been diagnosed (Years diag) with PD. Values represent mean (SE).

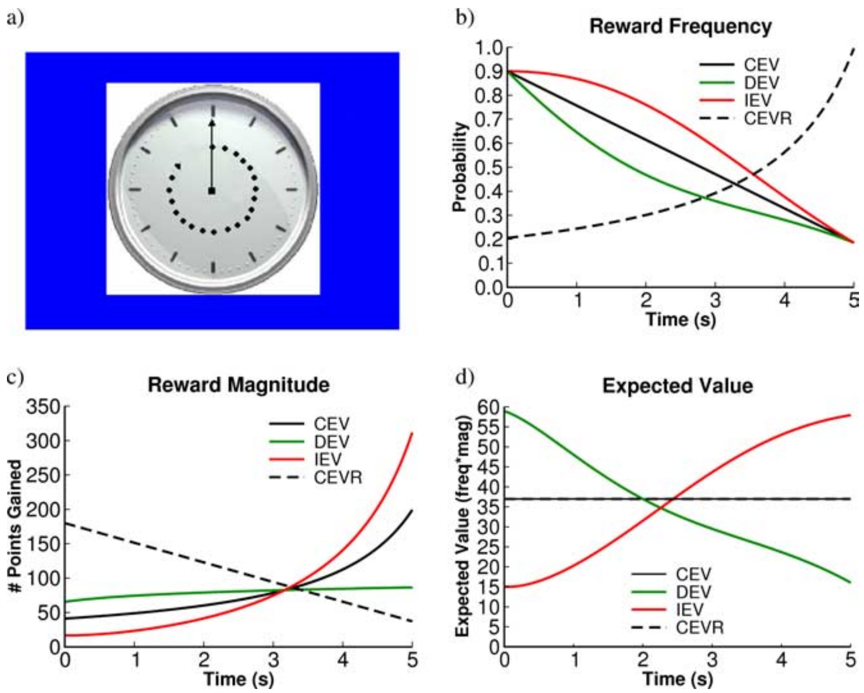

Figure 1. Task conditions: DEV, CEV, IEV, and CEVR. The *x*-axis in all plots corresponds to the time after onset of the clock stimulus at which the response is made. The functions are designed such that the expected value in the beginning in DEV is approximately equal to that at the end in IEV so that if optimal, subjects should obtain the same average reward in both IEV and DEV. ***a***, Example clock-face stimulus; ***b***, probability of reward occurring as a function of response time; ***c***, reward magnitude (contingent on ***a***); ***d***, expected value across trials for each time point. Note that CEV and CEVR have the same EV, so the black line represents EV for both conditions.

the opposite bias (Frank et al., 2004). Similar results have since been observed as a result of DA manipulations in other populations and tasks (Cools et al., 2006; Frank and O'Reilly, 2006; Pessiglione et al., 2006; Frank et al., 2007c; Shohamy et al., 2008). Here, we examined whether the same theoretical framework can apply to reward maximization by response time adaptation. Our computational model predicts that striatal DA supports response speeding to maximize rewards as a result of positive reward prediction errors, whereas low DA levels support response slowing caused by negative prediction errors. We test these predictions in a novel task which requires making only a single response. In addition to the main conditions of interest, the task also enabled us to study rapid trial-to-trial adjustments, and a bias to learn more about the frequency versus magnitude of rewards.

Materials and Methods

Sample. We tested 17 healthy controls and 20 Parkinson's patients both off and on medications (Table 1). Parkinson's patients were recruited from the University of Arizona Movement Disorders Clinic. The majority of patients were taking a mixture of dopaminergic precursors (levodopa-containing medications) and agonists. (Six patients were on DA agonists only and three patients on DA precursors only.) Control subjects were either spouses of patients (who tend to be fairly well matched demographically), or recruited from local Tucson senior centers.

Task. Participants were presented a clock face whose arm made a full turn over the course of 5 s. They were instructed as follows.

"*You will see a clock face. Its arm will make a full turn over the course of 5 s. Press the 'spacebar' key to win points before the arm makes a full turn. Try to win as many points as you can!*

"*Sometimes you will win lots of points and sometimes you will win less. The time at which you respond affects in some way the number of points that you can win. If you don't respond by the end of the clock cycle, you will not win any points.*

"*Hint: Try to respond at different times along the clock cycle to learn how to make the most points. Note: The length of the experiment is constant and is not affected by when you respond.*"

The trial ended after the subject made a response or if the 5 s duration elapsed and the subject did not make response. Another trial started after an intertrial interval (ITI) of 1 s.

There were four conditions, comprising 50 trials each, in which the probabilities and magnitudes of rewards varied as a function of time elapsed on the clock until the response. Before each new condition, participants were instructed: "*Next, you will see a new clock face. Try again to respond at different times along the clock cycle to learn how to make the most points with this clock face.*"

In the three primary conditions considered here (DEV, CEV, and IEV), the number of points (reward magnitude) increased, whereas the probability of receiving the reward decreased, over time within each trial. Feedback was provided on the screen in the form of "You win XX points!". The functions were designed such that the expected value (probability*magnitude) either decreased (DEV), increased (IEV), or remained constant (CEV), across the 5 s trial duration (Fig. 1). Thus in the DEV condition, faster responses yielded more points on average, whereas in the IEV condition slower responses yielded more points. [Note that despite high frequency of rewards during early periods of IEV, the small magnitude of these rewards relative to other conditions and to later responses would actually be associated with negative prediction errors (Holroyd et al., 2004; Tobler et al., 2005).] The CEV condition was included for a within-subject baseline RT measure for separate comparisons with IEV and DEV. In particular, because all response times are equivalently adaptive in the CEV condition, the participants' RT in that condition controls for potential overall effects of disease or medication on motor responding. Given this baseline RT, an ability to learn adaptively to integrate expected value across trials would be indicated by relatively faster responding in the DEV condition and slower responding in the IEV condition.

In addition to the above primary conditions, we also included another "CEVR" condition in which expected value is constant, but reward probability increases whereas magnitude decreases as time elapses (CEVR = CEV Reverse). This condition was included for multiple reasons. First, because both CEV and CEVR have equal expected values across all of time, any difference in RT in these two conditions can be attributed to a participant's potential bias to learn more about reward probability than about magnitude or vice versa. Specifically, if a participant waits longer in

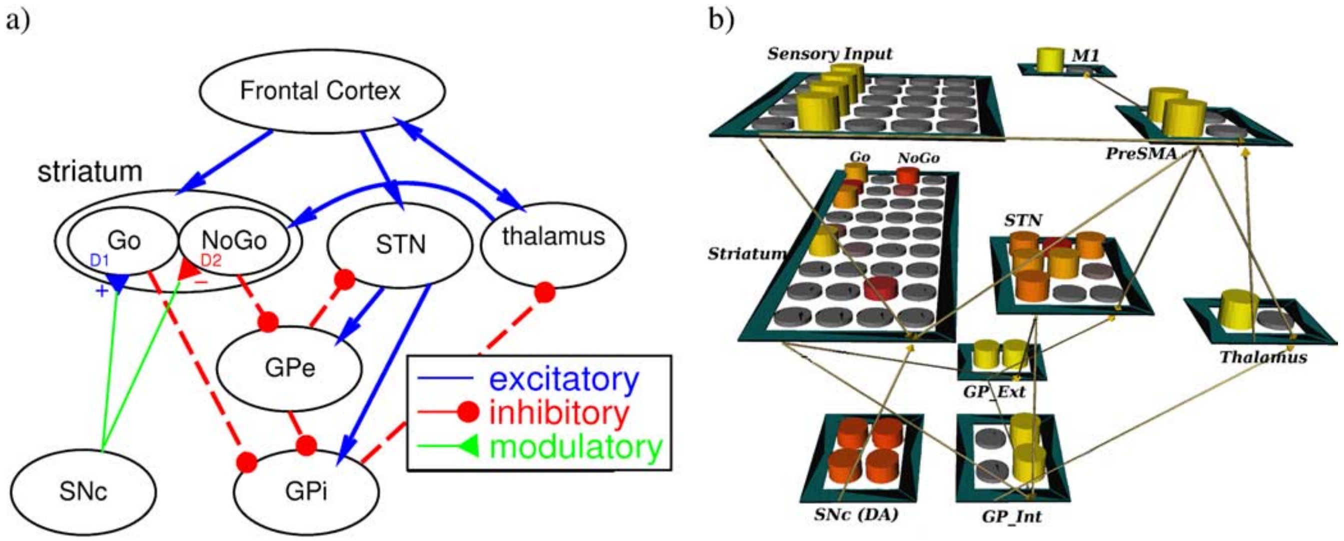

Figure 2. ***a***, Functional architecture of the model of the basal ganglia. The direct ("Go") pathway disinhibits the thalamus via the interior segment of the GPi and facilitates the execution of an action represented in the cortex. The indirect (NoGo) pathway has an opposing effect of inhibiting the thalamus and suppressing the execution of the action. These pathways are modulated by the activity of the substantia nigra pars compacta (SNc) that has dopaminergic projections to the striatum. Go neurons express excitatory D1 receptors whereas NoGo neurons express inhibitory D2 receptors. ***b***, The Frank (2006) computational model of the BG. Cylinders represent neurons, height and color represent normalized activity. The input neurons project directly to the pre-SMA in which a response is executed via excitatory projections to the output (M1) neurons. A given cortical response is facilitated by bottom-up activity from thalamus, which is only possible once a Go signal from striatum disinhibits the thalamus. The left half of the striatum are the Go neurons, the right half are the NoGo neurons, each with separate columns for responses R1 and R2. The relative difference between summed Go and NoGo population activity for a particular response determines the probability and speed at which that response is selected. Dopaminergic projections from the substantia nigra pars compacta (SNc) modulate Go and NoGo activity by exciting the Go neurons (D1) and inhibiting the NoGo neurons (D2) in the striatum, and also drive learning during phasic DA bursts and dips. Connections with the subthalamic nucleus (STN) are included here for consistency, and modulate the overall decision threshold Frank (2006), but are not relevant for the current study.

CEVR than in CEV, it can be inferred that the participant is risk averse because they value higher probabilities of reward more than higher magnitudes of reward. Moreover, despite the constant expected value in CEVR, if one is biased to learn more from negative prediction errors, they will tend to slow down in this condition because of the high probability of their occurrence. Finally, the CEVR condition also allows us to disentangle whether trial-to-trial RT adjustment effects reflect a tendency to change RTs in the same direction after gains, or whether RTs might change in opposite directions after rewards depending on the temporal envelope of the reward structure (see below).

The order of condition (CEV, DEV, IEV, CEVR) was counterbalanced across participants. A rest break was given between each of the conditions (after every 50 trials). Subjects were instructed in the beginning of each condition to respond at different times to try to win the most points, but were not told about the different rules (e.g., IEV, DEV). Each condition was also associated with a different color clock face to facilitate encoding that they were in a new context, with the assignment of condition to color counterbalanced.

To prevent participants from explicitly memorizing a particular value of reward feedback for a given response time, we also added a small amount of random uniform noise (±5 points) to the reward magnitudes for each trial; nevertheless the basic relationships depicted in Figure 1 remain.

Relation to other temporal choice paradigms. Most decision making paradigms study different aspects of which motor response to select, generally not focusing on temporal aspects of when responses are made. Perhaps the most relevant intertemporal choice paradigm is that of "delay discounting" (McClure et al., 2004, 2007; Hariri et al., 2006; Scheres et al., 2006; Heerey et al., 2007). Here, subjects are asked to choose between one option that leads to small immediate reward, versus another that would produce a large, but delayed reward. On the surface, these tasks bear some similarity to the current task, in that choices are made between different magnitudes of reward values that occur at different points in time. Nevertheless, the current task differs from delay discounting in several important respects. First, our task requires selection of only a single response, in which the choice itself is determined only by its latency, over the course of 5 s. In contrast, the delay discounting paradigm involves multiple responses for which latency of the reward differs, over longer time courses of minutes to weeks and even months, but in which the latency of the response itself is not relevant. Second, our task is less verbal, and more experiential. That is, in delay discounting, participants are explicitly told the reward contingencies and are simply asked to reveal their preference, trading off reward magnitude against the delay of its occurrence. In contrast, subjects in the current study must learn statistics of reward probability, magnitude, and their integration, as a result of experience across multiple trials within a given context. This process is likely implicit, a claim supported by the somewhat subtle (but reliable) RT adjustments in the task, together with informal analysis of postexperimental questionnaires in young, healthy pilot participants (and a subset of patients here), who showed no explicit knowledge of time-reinforcement contingencies.

Analysis. Response times were log transformed in all statistical analyses to meet statistical distributional assumptions (Judd and McClelland, 1989). For clarity, however, raw response times are used when presenting means and SEs. To measure learning within a given condition, we also compared response times in the first block of 12 trials within each condition (the first quarter) to that of the last block of 12 trials in that condition. Statistical comparisons were performed with SAS 9.1.3 proc MIXED to examine both between- and within-subject differences, using unstructured covariance matrices (which does not make any strong assumptions about the variance and correlation of the data, as do structured covariances).

Computational modeling. In addition to the empirical study, we also simulated the task using our computational neural network model of the basal ganglia (Frank, 2006), as well as a more abstract "temporal difference" (TD) simulation (Sutton and Barto, 1998). The neural model simulates systems-level interactive neural dynamics among corticostriatal circuits and their roles in action selection and reinforcement learning (Fig. 2). Neuronal dynamics are governed by coupled differential equations, and different model neurons for each of the simulated areas to capture differences in physiological and computational properties of the regions comprising this circuit. We refrain from reiterating all details of the model (including all equations, detailed connectivity, parameters, and their neurobiological justification) here; interested readers should refer to Frank (2006) and/or the on-line database *modelDB* wherein the previous simulations are available for download. The model can also be obtained by sending an E-mail to mfrank@u.arizona.edu. The same model parameters were used in previous human simulations in choice tasks, so that the simulation results can be considered a prediction from a priori modeling rather than a "fit" to new data.

Neural model high level summary. We first provide a concise summary here of the higher level principles governing the network functionality, focusing on aspects of particular relevance for the current study. Two separate "Go" and "NoGo" populations within the striatum learn to facilitate and suppress responses, with their relative difference in activity states determining both the likelihood and speed at which a particular response is facilitated. Separately, a "critic" learns to compute the expected value of the current stimulus context, and actual outcomes are computed as prediction errors relative to this expected value. These prediction errors train the value learning system itself to improve its predictions, but also drive learning in the Go and NoGo neuronal populations. As positive reward prediction errors accumulate, phasic DA bursts drive Go learning via simulated D1 receptors, leading to incrementally speeded responding. Conversely, an accumulation of negative prediction errors encoded by phasic DA dips drive NoGo learning via D2 receptors, leading to incrementally slowed responses. Thus, a sufficiently high dopaminergic response to a preponderance positive reward prediction errors is associated with speeded responses across trials, but sufficiently low striatal DA levels are necessary to slow responses because of a preponderance of negative prediction errors.

Connectivity. The input layer represents stimulus sensory input, and projects directly to both premotor cortex (e.g., pre-SMA) and striatum. Premotor units represent highly abstracted versions of all potential responses that can be activated in the current task. However, direct input to premotor activation is generally insufficient in and of itself to execute a response (particularly before stimulus-response mappings have been ingrained). Rather, coincident bottom-up input from the thalamus is required to selectively facilitate a given response. Because the thalamus is under normal conditions tonically inhibited by the globus pallidus (basal ganglia output), responses are prevented until the striatum gates their execution, ultimately by disinhibiting the thalamus.

Action selection. To decide which response to select, the striatum has separate "Go" and "NoGo" neuronal populations that reflect striatonigral and striatopallidal cells, respectively. Each potential cortical response is represented by two columns of Go and NoGo units. The globus pallidus nuclei effectively compute the striatal Go − NoGo difference for each response in parallel. That is, Go signals from the striatum directly inhibit the corresponding column of the globus pallidus (GPi). In parallel, striatal NoGo signals inhibit the GPe (external segment), which in turn inhibits the GPi. Thus a strong Go − NoGo striatal activation difference for a given response will lead to a robust pause in activity in the corresponding column of GPi, thereby disinhibiting the thalamus and allowing bidirectional thalamocortical reverberations to facilitate a cortical response. The particular response selected will generally be the one with the greatest Go − NoGo activity difference, because the corresponding column of GPi units will be most likely and most quickly inhibited, allowing that response to surpass threshold. Once a given cortical response is facilitated, lateral inhibitory dynamics within cortex allows the other competing responses to be suppressed.

Note that the relative Go-NoGo activity can affect both which response is selected, and also the speed with which it is selected. [In addition, the subthalamic nucleus can also dynamically modify the overall response threshold, and therefore response time, in a given trial by sending diffuse excitatory projections to the GPi (Frank, 2006). This functionality enables the model to be more adept at selecting the best response where there is high conflict between multiple responses, but is orthogonal to the point studied here, so we do not discuss it further.]

Learning attributable to DA bursts and dips. How do particular responses come to have stronger Go or NoGo representations? Dopamine from the substantia nigra pars compacta modulates the relative balance of Go versus NoGo activity via simulated D1 and D2 receptors in the striatum. This differential effect of DA on Go and NoGo units, via D1 and D2 receptors, affects performance (i.e., higher levels of tonic DA leads to overall more Go and therefore a lower threshold for facilitating motor responses and faster RTs), and critically, learning. Phasic DA bursts that occur during unexpected rewards drive Go learning via D1 receptors, whereas phasic DA dips that occur during unexpected reward omissions drive NoGo learning via D2 receptors. These dual Go/NoGo learning mechanisms proposed by our model (Frank, 2005) are supported by recent synaptic plasticity studies in rodents (Shen et al., 2008).

Because there has been some question of whether DA dips confer a strong enough signal to drive negative prediction errors, we outline here a physiologically plausible account based on our modeling framework (Frank and Claus, 2006). Importantly, D2 receptors in the high-affinity state are much more sensitive than D1 receptors (which require significant bursts of DA to get activated). This means that D2 receptors are inhibited by low levels of tonic DA, and that the NoGo learning signal depends on the extent to which DA is removed from the synapse during DA dips. Notably, larger negative prediction errors are associated with longer DA pause durations of up to 400 ms (Bayer et al., 2007), and the half-life of DA in the striatal synapse is 55–75 ms (Gonon, 1997; Venton et al., 2003). Thus, the longer the DA pause, the greater likelihood that a particular NoGo-D2 unit would be disinhibited, and the greater the learning signal across a population of units. Furthermore, depleted DA levels as in PD would enhance this effect, because of D2 receptor sensitivity (Seeman, 2008) and enhanced excitability of striatopallidal NoGo cells in the DA-depleted state (Surmeier et al., 2007).

To foreshadow the simulation results, responses that have had a larger number of bursts than dips in the past will therefore have developed greater Go than NoGo representations and will be more likely to be selected earlier in time. Early responses that are paired with positive prediction errors will be potentiated by DA bursts and lead to speeded RTs (as in the DEV condition), whereas those responses leading to less than average expected value (negative prediction errors) will result in NoGo learning and therefore slowing (as in the IEV condition). Moreover, manipulation of tonic and phasic dopamine function (as a result of PD and medications) should then affect Go vs NoGo learning and associated response times.

Model methods for the current study. We include as few new assumptions as possible in the current simulations. The input clock-face stimulus was simulated by activating a set of four input units representing the features of the clock – this was the same abstract input representation used in other simulations. Because our most basic model architecture includes two potential output responses [but see Frank (2006) for a four-alternative choice model], we simply added a strong input bias weight of 0.8 to the left column of premotor response units. The exact value of this bias is not critical; it simply ensures that when presented with the input stimulus in this task, the model would always respond with only one response (akin to the spacebar in the human task), albeit at different potential time points. Thus this input bias weight is effectively an abstract representation of task-set.

Models were then trained for 50 trials in each of the conditions (CEV, IEV, DEV, CEVR) in which reward probability and magnitude varied in an analogous manner to the human experiments. The equations governing probability, magnitude and expected value were identical to those in the experiment, as depicted in Figure 1. We also had to rescale the reward magnitudes from the actual task to convert to DA firing rates in the model, and to rescale time from seconds to units of time within the model, measured in processing cycles.

Specifically, reward magnitudes were rescaled to be normalized between 0 and 1, and the resulting values applied to the dopaminergic unit phasic firing rate during experienced rewards. A lack of reward was simulated with a DA dip (no firing), and maximum reward is associated with maximal DA burst. Furthermore, because phasic DA values are scaled in proportion to reward magnitude, only relatively large rewards were associated with an actual DA burst that is greater than the tonic value. Rewards smaller than expected value lead to effective DA dips. This function was implemented by initializing expected value at the beginning of each condition to zero, and then updating this value according to the standard Rescorla-Wagner δ rule: $V(t+1) = V(t) + \alpha(R - V(t))$, where α is a learning rate for integrating expected value and was set to 0.1. Thus, as the model experienced rewards in the task, subsequent rewards were encoded relative to this expected value and then applied to the phasic DA firing rate. Together these features are meant to reflect the observed property of DA neurons in monkeys, where phasic firing is proportional to reward prediction error rather than reward value itself, and where firing rates are normalized relative to the largest current available reward (Tobler et al., 2005). Similar relative prediction error encoding has been observed in humans using electrophysiological measures thought to be related to phasic DA signaling (Holroyd et al., 2004). The implication in the current task is that, for example, in the IEV condition, although reward frequency is highest early in the trial, the magnitude of rewards is lower than average in this period, and therefore should be associated with a "dip" in DA to encode the low expected value of this period.

Time was rescaled from a maximum of 5 s of real time to a maximum of 200 cycles of processing in the model, where each cycle reflects one update of neuronal membrane potentials as a function of their updated weighted inputs and subject to time constants of ionic channel activation and membrane capacitance. The model response time was measured as in previous work (Frank et al., 2007b,d) (T. V. Wiecki, K. Riedinger, A. Meyerhofer, W. J. Schmidt, and M. J. Frank, unpublished observations). Specifically, as described above, the basal ganglia select responses by removing tonic inhibition onto the thalamus (Mink, 1996; Chevalier and Deniau, 1990). Thus a response is selected when the associated thalamus units' activity exceeds 50% maximal firing. (Because the BG gating of thalamic activity is required to facilitate a cortical response, similar results are obtained by probing output unit activity, however the thalamic activity is a more direct assessment of BG output, given that an output unit can still sometimes fire noisily).

Parkinson's disease and DA medications were also simulated as reported previously (Frank et al., 2004, 2007b; Frank, 2005). To simulate PD, we reduced the number of intact DA units from 4 to 2, such that overall DA activity, both tonic and phasic, was reduced. To simulate DA medications, DA activity was restored but prevented from decreasing all the way to zero during DA dips. That is, there was a non-zero minimum value of DA activity, simulating the tonic stimulation of DA agonist medication onto D2 receptors even during potential pauses in actual DA unit firing (Frank et al., 2004; Frank, 2005). This effectively impairs networks from learning NoGo to non-rewarded responses.

Temporal difference model. We also examined whether a standard temporal difference (TD[λ]) reinforcement learning model (Montague et al., 1996; Schultz et al., 1997; Sutton and Barto, 1998; McClure et al., 2003) could capture the pattern of behavior observed in the humans and neural network model. In this model, the TD reward prediction error δ in trial t at time point i is computed according to current rewards plus summed discounted future rewards relative to the current value estimate: $\delta_{t,i} = r_{t,i} + \gamma V_{t,i+1} - V_{t,i}$, where r is reward value, γ is a discount parameter, and V is the expected value at each time point. Then the value $V_{t+1,i}$ is updated according to $V_{t,i} + \alpha\delta_{t,i}e_i$, where i is the time step within each trial (we used 10 time steps) α is a learning rate, and e_i is the eligibility trace of the stimulus representation x_i. This eligibility trace is calculated based on a "complete serial compound" representation of the stimulus, in which a different x_i represents a stimulus at each point in time, and the weights for these x_i remain eligible to learn for an exponentially decaying period thereafter. Specifically, the eligibility trace is updated at each time step: $e_i = \lambda\gamma e_{i-1} + x_i$, where λ is the trace decay parameter. [Eligibility traces are also scaled by γ because the credit assigned to previous time points must be discounted by the relatively longer wait to future reward (Sutton and Barto, 1998).] This TD[λ] implementation allows the algorithm to immediately assign reward credit to events that occurred in the past, and in effect causes temporal representations to be "smeared" such that reward values occurring at particular times are generalized to a range of earlier time points.

These equations should allow the algorithm to learn the reward values of the clock face stimulus at different points in time. To generate response times, we follow McClure et al. (2003), who captured aspects of incentive motivation literature using TD by assuming that response times are a function of reward prediction error at any given time [based on previous work by Egelman et al. (1998)]. Specifically, at each time point during the trial, the model generates a probability of responding $P = (1/(1 + e^{-m(\delta i-b)}))$, where m is a scaling constant and was set to 0.8, and b affects the base-rate probability and was set to 2 (such that on average with $\delta = 0$, the model responds halfway through the trial, as in the BG model. When the model makes a response, a probability and magnitude of reward is calculated according to the equations used for the subjects and neural network model described above. Response times are scaled such that each time step in the model corresponds to 500 ms for humans.

We ran simulations using all combinations of γ, α and λ from 0.1–1 in steps of 0.1. Each point in parameter space was simulated 20 times, each time initializing the weights (i.e., 20 different "subjects"). Our analysis approach was to select parameters based on the model's performance in two control tasks. In the first control task, the model was given a reward at time = 30 on each trial (equal magnitude on each trial), and was rewarded in different conditions with probability of 25%, 50%, 75%, or 100%. The idea is that the model should respond earlier during conditions with higher probability of reward, because value increases and propagates backwards in time to the onset of the stimulus. The second control task varies the magnitude of reward as well as probability and tests whether the RT is modulated by expected value, given that this is a requirement of the experimental task. The model performed both of these control tasks well at most parameter settings. The following average of parameters produced a linear decrease in response time with increasing reward probability: $\gamma = 0.6–0.8$; $\lambda = 0.6–0.7$; $\alpha = 0.3–0.4$. We then plotted the responses of the model during the experimental task using these same parameters, but also searched a range of other parameter sets.

Table 2. Response times (milliseconds) in each task condition for each group, across all trials

Block/group	DEV	IEV	CEV	CEVR
Seniors	1697 (142)	2211 (136)	1988 (122)	2516 (119)
PD off medication	1831 (152)	2393 (235)	1940 (196)	2148 (154)
PD on medication	1785 (139)	2244 (150)	1967 (114)	1995 (123)

Values reflect mean (SE).

Results

We predicted that increases in striatal DA in medicated PD should enhance Go learning but impair NoGo learning, because of blockade of DA dips needed to learn NoGo (Frank et al., 2004; Frank, 2005). In contrast, depleted striatal DA should potentiate NoGo learning at the expense of reduced Go learning. Thus we predicted that patients on medication would show relatively speeded RTs in the DEV condition, but would not slow down in IEV, whereas patients off medication would show the opposite pattern. We further predicted that any trial-to-trial adjustments in RT should not be related to BG DA levels (Frank et al., 2007a), but might instead reflect integrity of prefrontal cortex and associated dopamine innervation.

Table 2 shows the mean RT in each condition (across all trials), and Figure 3 shows RTs as a function of trial number in each condition. There were no differences between patients and controls, or medication effects, on baseline RTs in the CEV condition, or on overall response time (p values >0.3). Overall, across all subjects, response times in the IEV condition were significantly higher than those in the DEV condition ($t_{(36)} = 4.72$, $p < 0.0001$). Thus, although participants were not optimal, which would require responding immediately in DEV and waiting until just before 5 s elapsed in IEV, they nevertheless learned to adapt RTs in the direction expected. However, any differences between DEV and IEV themselves could be attributed to speeding (Go learning) in DEV, or slowing (NoGo learning) in IEV, or both. Thus, the main measures of interest are DEV (within-subject speeding in DEV relative to baseline CEV) and IEV (within-

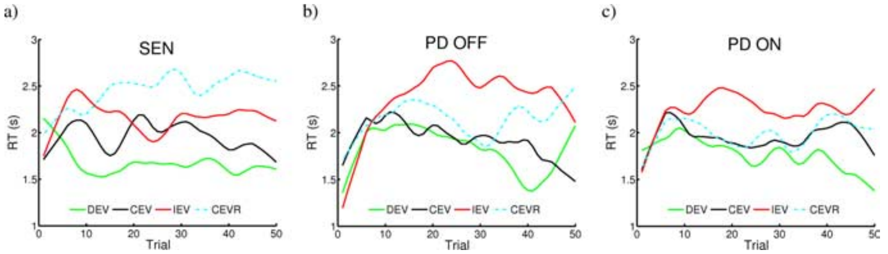

Figure 3. ***a–c***, Response times as a function of trial number, smoothed with a 10 trial kernel, in healthy seniors (***a***), patients off medication (***b***), and patients on medication (***c***).

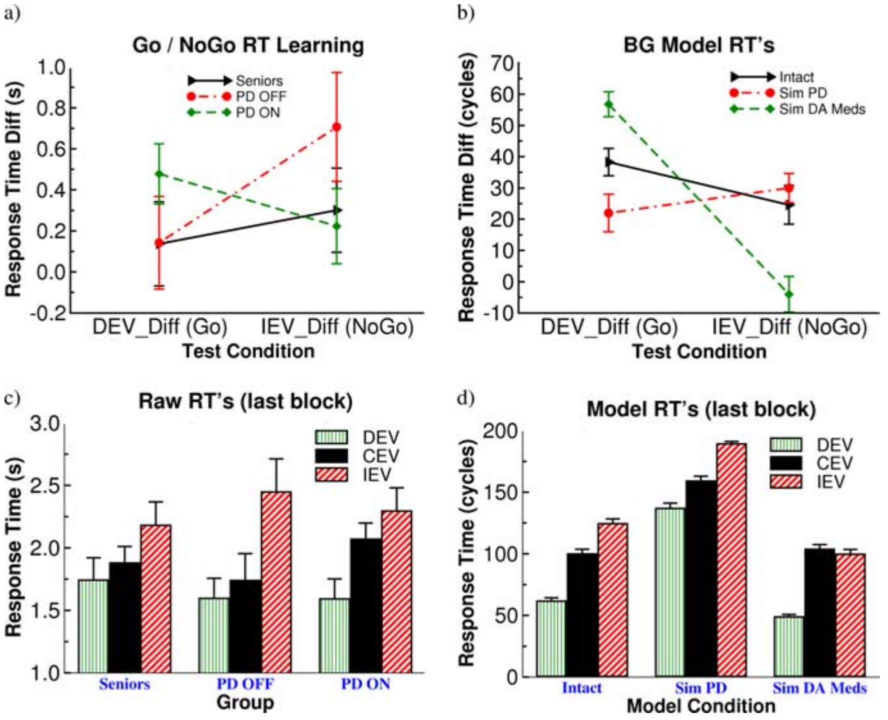

Figure 4. ***a***, Relative within-subject biases to speed RTs in DEV compared with CEV (Go learning) and to slow RTs in IEV compared with CEV (NoGo learning). Values represent mean (SE) in the last block of 12 trials in each condition. ***b***, Similar pattern of results from neural network model of the basal ganglia. ***c***, ***d***, Raw response times are shown for each condition in participants in this study (***c***) the neural model (***d***) (see Materials and Methods, Model methods for the current study, for quantification of model RTs).

subject slowing in IEV relative to baseline CEV), which are measures of Go and NoGo learning, respectively.

As predicted, while on medication, PD patients were better able to speed up in the DEV condition, as measured by their RT change from the first to last block. The interaction between medication status (off vs on) and block in the DEV condition was significant ($F_{(1,35)}$ = 5.63, p = 0.02). Conversely, the opposite pattern of results was observed in IEV: while off medication, patients were better able to slow their responses across blocks ($F_{(1,35)}$ = 6.7, p = 0.02). Moreover, a within-subject analysis of the difference between Go learning (DEV) and NoGo learning (IEV) conditions reveals a significant interaction between medication status and block ($F_{(1,35)}$ = 5.25, p = 0.03). Neither patients on or off medication differed significantly from age-matched controls in DEV or IEV (p values >0.2). Further, controls' performance did not differ between the first and last blocks in DEV (Go) ($F_{(1,35)}$ = 0.02), IEV(NoGo) ($F_{(1,35)}$ = 0.29), or the difference between DEV and IEV ($F_{(1,35)}$ = 0.18) conditions. Nevertheless, the lack of learning effect across blocks in healthy controls reflects the fact that they learned early within the first block, whereas patients required more training trials to differentiate between the conditions (Fig. 3). Indeed, when measured across all trials, healthy seniors showed learning in both conditions, as reflected by DEV and IEV which were both significantly greater than zero (p <= 0.05 for both comparisons). This pattern of results confirms the above mentioned hypothesis that speeding or slowing down RT is respectively enhanced by increases or decreases in striatal DA levels (Frank et al., 2004; Cools et al., 2006; Frank and O'Reilly, 2006).

We also analyzed RT performance in the last block by itself (which should reflect stabilized learning). Mirroring the above significant findings across blocks, there was marginal interaction between DEV and IEV and medication status ($F_{(1,35)}$ = 3.6, p = 0.06), such that the on medication state was associated with relatively better Go learning but worse NoGo learning (Fig. 4a).

Finally, we found that DEV negatively correlated with IEV across all participants (r(36) = −0.33, p = 0.02). This correlation was significant in PD patients alone (r(19) = −0.4, p = 0.02) but not in control subjects (r(16) = −0.17, p = 0.47), suggesting that this negative correlation is accentuated by having DA levels restricted to the low or high end. This finding supports the hypothesis that the same mechanism that leads to adaptive DEV responding causes impairment in IEV, and vice versa. Overall, these results also confirm the hypothesis that DA manipulations modulate Go and NoGo learning in opposite directions (Fig. 4a), a finding that was captured by the BG neural network model (Fig. 4b), but not the temporal difference simulations (Fig. 5) (see below for detailed analysis and discussion).

Probability-magnitude bias

We also analyzed the difference in RT between CEV and CEVR conditions as a measure of probability-magnitude bias (PM-bias = CEVR − CEV). We found PM-bias to be positive across all groups (Fig. 6), indicating an avoidance of choices associated with low reward probability, consistent with risk aversion behavior (Kahneman and Tversky, 1979). We also found medicated patients to be less risk averse in this sense than control subjects ($F_{(1,35)}$ = 5.46, p = 0.02), with a similar trend when compared with their nonmedicated state ($F_{(1,35)}$ = 3.12, p = 0.08). These results are consistent with those described above, in that "risk aversion" in this context depends on learning in CEVR that rewards are improbable for early responses and to therefore slow down. Thus the observation that medicated patients show less of a PM-bias is consistent with their impaired NoGo learning, as evidenced in the IEV condition as well.

Trial-to-trial RT adaptation

Finally, to further investigate the source of the learning biases, we analyzed trial-to-trial changes in RT as a function of whether the previous trial led to a reward or not. Across all subjects, we found an effect of feedback type on RT adaptation. In the DEV, CEV, and IEV conditions, subjects had significantly longer RTs after receiving positive than negative feedback (all p values <0.02) (Fig. 7). The opposite pattern held true for the CEVR condition ($t_{(36)} = -2.09$, $p = 0.04$). Could these trial-to-trial effects form the basis for the learning biases above? Notably, there was no effect of medication on this RT adaptation from single rewards or the lack thereof, in any task condition (all comparisons insignificant, and all but one $p > 0.3$; Fig. 7). Critically, the trial-to-trial adaptations were for the most part in the opposite direction to the cumulative RT changes: in the three primary conditions (CEV, DEV, IEV) participants actually slowed down after wins, and sped up after non-rewards. These contrast with the incremental adjustments in which conditions associated with high gains overall early in time were associated with response speeding (Go learning in DEV) across trials, whereas conditions associated with low gains early in trial were associated with response slowing (NoGo learning in IEV).

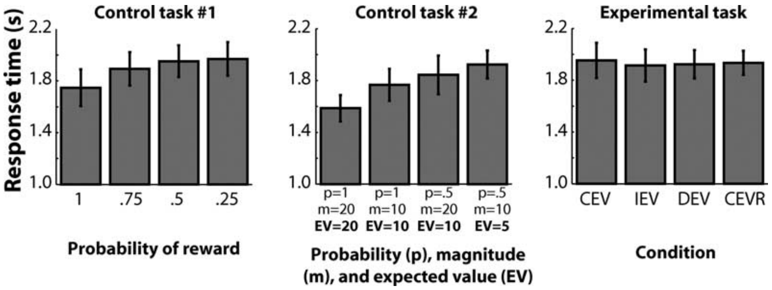

Figure 5. Temporal difference model results. Control task 1, Control task showing that the TD implementation can successfully speed responses for stimuli that have a greater probability of being followed by a reward with constant delay (i.e., showing Pavlovian to instrumental transfer) across a range of parameters. Control task 2, Similar results for increasing expected value. Experimental task, The same TD model fails to differentially modulate RTs across conditions within our experimental task. See Results, Temporal difference simulation results.

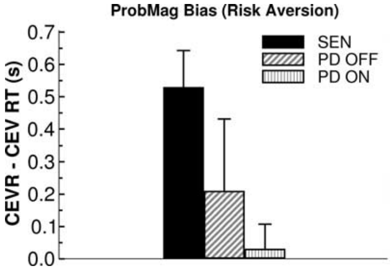

Figure 6. Relative within-subject biases to prefer high probability over high magnitude, controlling for equal expected value (CEVR − CEV). Senior controls and patients off medication showed risk aversion, whereas those on medication did not. Values represent mean (SE) in the last block of trials in each condition.

At first glance, these findings appear that participants change their RTs such that after wins, they are more willing to risk a low probability of reward to potentially win a larger gain, and after reward omissions, they become more conservative. However, further scrutiny of the data reveals that these effects are entirely the result of the fact that participants tend to speed up after slow responses and slow down after fast responses, as they explore the space of reward structure for different response times. Figure 7*d* shows RT adjustments across all conditions after responses that are faster or slower than the mean response. These effects are of much larger magnitude than those conditionalized on previous rewards, described above. Indeed, the apparent differential adjustment after rewards and lack thereof is an artifact of a sampling bias: in the three primary conditions, gains are more frequent after faster responses (by design), and losses are more frequent after slower responses; this relationship reverses in CEVR. When RT analysis is confined to a particular range (e.g., within a SD of the mean), the differences resulting from previous rewards disappear. Thus the rapid trial-to-trial adjustments seem to reflect participants' explicit tendency to explore the space to determine the reward structure. A more explicit investigation of these effects, along with a mathematical model that attempts to rationalize exploratory tendencies based on the uncertainty of reward/response contingencies, will be presented in future work (M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno, unpublished observations).

Basal ganglia model simulation results

The finding that dopaminergic medication status affects incremental Go vs NoGo RT adaptation provides converging confirmatory support for a computational model of basal ganglia function (Fig. 2), hitherto applied to simulate DA medication effects on various other reinforcement learning phenomena in both humans (Frank et al., 2004, 2007b; Frank, 2005) and rats (Wiecki, Riedinger, Meyerhofer, Schmidt, and Frank, unpublished observations). Here we show explicitly that this same model captures the qualitative pattern of results found in this study (see above for methods).

All reported results are averaged across 25 networks with different sets of initial random synaptic weights in each model condition (intact, PD, medication). As can be seen in Figure 4*b*, simulated DA manipulation mirrored that seen in PD patients on and off medication, albeit in a somewhat more idealized form. Simulated Parkinson networks were impaired at speeding up in the DEV condition relative to their CEV baseline RT, but showed enhanced ability to slow down in the IEV condition. This is because low DA levels potentiated NoGo learning, which in turn led to enhanced IEV slowing (see also Wiecki, Riedinger, Meyerhofer, Schmidt, and Frank, unpublished observations).

In direct contrast, simulated DA medication restored speeding in the DEV condition but led to an inability to slow down in IEV, because of an effective blockade of DA dips needed to learn NoGo – and therefore an inability to learn that the small magnitude rewards are actually "worth" less than expected. Finally, intact networks showed some degree of both Go and NoGo learning. Although the pattern in this case was such that Go learning was somewhat potentiated relative to NoGo learning, which was not seen in healthy seniors, we note that the exact quantitative data are not critical here (given that we did fit the model to the data), but rather the qualitative pattern of data as a function of DA manipulation.

Finally, we also ran the model in the CEVR condition, and found that in the intact case, there was no significant difference between RT in the last block of the CEV condition, showing no probability or magnitude bias (mean PM-bias = −6.3 cycles, SE = 5.3). However, simulated PD networks showed longer RTs

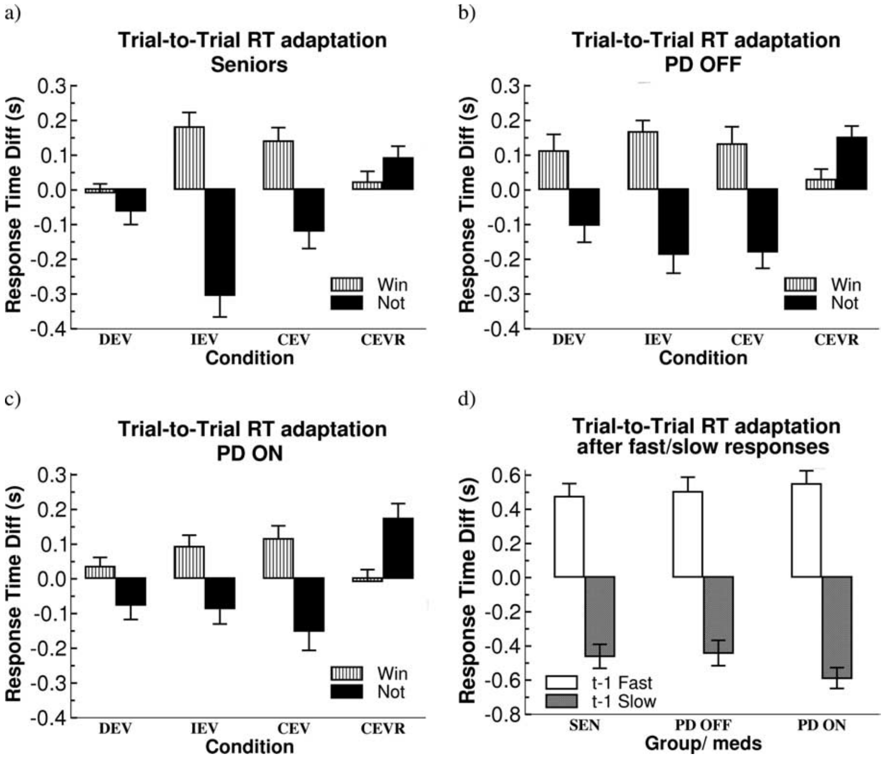

Figure 7. ***a–d***, Trial-to-trial adjustments in RT from previous to current trial, conditionalized according to whether the last trial was rewarded (Win) or not in senior controls (***a***), PD patients off medication (***b***), and patients on medication (***c***). ***d***, Trial-to-trial adjustments across all conditions after faster and slower than average responses. Note difference in scale. Values represent mean (SE).

in CEVR than CEV (mean PM-bias = 22 cycles, SE = 4.8), whereas medicated networks showed the reverse pattern (mean PM-bias = −20.8 cycles, SE = 4.8). Thus, although networks on average did not show risk aversion (greater propensity to seek high probability rewards), the model nevertheless captured the relatively greater tendency for nonmedicated than medicated patients to do so. This result likely occurred because the simulated PD networks have a NoGo learning bias, causing them to avoid responding early in time in the CEVR condition when rewards are very infrequent, together with reduced representation of reward magnitude caused by depleted phasic DA levels. In contrast, medicated networks were not "averse" to responding early in CEVR, because of NoGo learning deficits, and their enhanced (relative to PD networks) phasic DA bursts reinforced these early responses during large magnitude rewards.

In sum, the BG model provides an account for (1) increased DEV, (2) decreased IEV, and (3) decreased PM-bias in medicated relative to nonmedicated patients.

Temporal difference simulation results

The TD model does not inherently include a means to capture response times. However, as described in Materials and Methods, previous work showed that some aspects of incentive motivation can be captured by assuming that the probability of a response being executed scales with the TD prediction error at that point in time (Egelman et al., 1998; McClure et al., 2003). Here we employ that model and show that it does successfully respond faster to stimuli that have a greater probability of being followed by reward, or greater expected value (both of which are associated with enhanced prediction errors at stimulus onset). However, when trained with the current temporal decision making task, the same model does not successfully learn to speed up RTs in the DEV condition or to slow down in IEV relative to CEV baseline (Fig. 5). Results shown are for the same parameter range producing satisfactory results for the control task, although we investigated the full range of parameters and still did not find a range of parameters that reliably produced the correct pattern of RTs. One reason for this failure is likely that TD propagates reward value back to the earliest predictor of its occurrence. In IEV, rewards are high late in the trial, but only if a response is made later. A problem with this is that as reward value is propagated to its earliest predictor, responses will then become faster, leading to lowered reward value, and causing instability. This issue is reminiscent of a failure of standard TD to maximize rewards in certain challenging RL tasks as a result of melioration (Montague and Berns, 2002). It is also likely that the variability in reward timing in the task (which are dependent on when responses are made), also posed somewhat of a challenge, given the known issues with timing variability in standard TD (Daw et al., 2003; O'Reilly et al., 2007).

Nevertheless, these simulation results certainly do not discomfirm or falsify the TD model. Our implementation assumes a very specific transformation of prediction errors to produce response times (that of McClure et al., 2003), and it is entirely possible that others, with additional assumptions, might learn the task appropriately (Niv et al., 2007) (see Discussion). Thus these simulations simply show that the standard TD model does not obviously capture results from this task (let alone account for effects of DA manipulation), in contrast to the neural model. It is not the added complexity of the neural model that allows it to capture these findings, as an abstract (reduced) mathematical implementation of our neural model can capture the RT adaptations by including a simple Go learning mechanism that accumulates positive prediction errors to drive RT speeding and a NoGo learning mechanism that accumulates negative prediction errors to drive RT slowing (Frank, Doll, Oas-Terpstra, and Moreno, unpublished observations).

Discussion

We showed that Parkinson's patients' tendency to adapt response times to maximize expected reward value depends on dopaminergic medication status. While off medication, patients tended to slow their responses to avoid early low expected values, but were less able to speed up when their early responses were rewarded. The opposite pattern was observed when the same patients were on dopaminergic medication; patients showed better response speeding, and worse response slowing, to maximize expected value. These results cannot be explained by overall differences in motor responding, because (1) there were no effects of medication status on overall RT in this task, and (2) the Go and NoGo learning measures were computed with respect to each individual's baseline "default" CEV response time. Moreover the same pattern of results was observed in our a priori model of reinforcement learning in the basal ganglia.

Our analysis rests on the idea that accumulated positive reward prediction errors implicitly drive basal ganglia-dependent Go learning to speed responses, whereas negative prediction errors drive NoGo learning and slowing. In choice paradigms, DA medications potentiate Go learning while also impairing NoGo learning (Frank et al., 2004, 2007b; Frank, 2005; Cools et al., 2006; Frank and O'Reilly, 2006; Pessiglione et al., 2006; Shohamy et al., 2008). Thus, the increased DEV-related speeding in medicated patients may reflect a combination of enhanced Go learning from positive feedback and reduced NoGo learning from omitted rewards (which would normally cause slowing). Similarly, increased IEV-related slowing in nonmedicated patients may reflect a combination of enhanced NoGo learning after outcomes that are worse than expected, and reduced Go learning after large positive outcomes.

This account is in accord with evidence from rodents and non-human primates that basal ganglia dopamine acts to speed responding when faced with rewarding cues (Satoh et al., 2003; Berridge, 2007; Niv et al., 2007), a process that is likely reliant on D1-dependent Go function. For example, the tendencies to speed responses to obtain large rewards and to approach a reward-predicting stimulus are both dependent on striatal D1 receptors (Dalley et al., 2005; Everitt and Robbins, 2005; Nakamura and Hikosaka, 2006). Conversely, striatal D2 receptor antagonism produces slowed responding when faced with lower than average rewards (Nakamura and Hikosaka, 2006). These D1 and D2 data converge with a recently reported study showing that these receptor types modulate synaptic plasticity in striatal Go and NoGo populations (Shen et al., 2008), and with human genetic data examining polymorphisms within dopaminergic genes and their effects on Go and NoGo learning (Frank et al., 2007a).

We note that in principle, speeded DEV responses could arise either from positive feedback/Go learning (as emphasized here), and/or participants could explicitly decide to change strategies after realizing that slowed responses produce bad outcomes. To the extent that such strategies are used in this task, they might be supported by other rule-based brain systems (e.g., Ashby and O'Brien, 2005; Daw et al., 2005). As far as the implicit basal ganglia reinforcement learning system is concerned, positive prediction errors accrued over multiple trials should serve to speed responses to associated stimuli, whereas negative prediction errors should slow responding. These incremental RT changes across trials lead to subtle differences between conditions (in the order of 200 ms), despite explicit decisions to respond early or late in any given trial.

Although the abstract TD model failed to produce the correct pattern of results between conditions, other variants might well be able to do so. For example, Niv et al. (2007) presented a model in which average reward rate, posited to be reported by tonic DA levels, served to increase vigor of responding. The formulation assumed that longer response latencies are associated with greater opportunity cost (caused by missed potential for rewards during the waiting period) and that this cost is directly proportional to the average reward rate. This model successfully captured observations that rats in free-operant tasks increase their response rates in proportion to the average reward rate. The authors note one possible way in which tonic DA levels come to represent average reward is via temporal integration/accumulation of phasic DA signals. With this assumption, the Niv et al. (2007) model may also capture the data reported here. Indeed, our BG model does so precisely because the DEV condition is associated with early positive prediction errors, phasic DA bursts, Go learning, and ultimately faster responding (and vice versa for IEV and negative prediction errors). Thus modulation of Niv's tonic DA/reward rate parameter to simulate PD and DA medications could potentially account for the cross-over-interactions observed here. Such an account can explain the clinical observation of bradykinesia in Parkinson's disease in motivational terms, whereby movement incurs a larger motivational cost because of lowered effective reward rate (Mazzoni et al., 2007; Niv and Rivlin-Etzion, 2007). In our model, this effect arises because of relatively increased activation and synaptic potentiation of striatopallidal "NoGo" neurons in the DA-depleted state (Surmeier et al., 2007; Shen et al., 2008).

Trial-to-trial adaptation

In addition to the incremental RT changes across trials, we also found evidence for rapid trial-to-trial RT adaptation that were in the opposite direction. Specifically, in the three primary conditions, participants tended to slow down after gains and speed up after losses. However, these effects were not affected by disease or medication status, and are therefore likely to rely on distinct neural systems. Further analysis revealed that these effects are likely to reflect trial-by-trial exploration, whereby fast responses are followed by slower responses and vice versa, as participants sample the probabilistic reward structure of each clock-face. In a recent genetic study, we reported that trial-to-trial adaptations were predicted by genetic factors controlling prefrontal dopaminergic function, whereas incremental probabilistic learning was predicted by genetic factors controlling striatal DA measures (Frank et al., 2007a). According to the neural models (Maddox and Filoteo, 2001; Ashby and OBrien, 2005; Frank and Claus, 2006), trial-to-trial adaptations rely on active maintenance of recent outcomes in the orbitofrontal cortex, whereas probabilistic learning depends on incremental synaptic weight changes in the basal ganglia (Knowlton et al., 1996; Packard and McGaugh, 1996; Graybiel, 1998, 2004; Delgado et al., 2000, 2005). However, our existing neural models of PFC function do not account for the trial-to-trial adaptations seen here, which may be more exploratory in nature. Indeed, we recently collected genetic data from 70 young healthy participants in this task, and found that prefrontal genetic function was associated with trial-to-trial exploration, whereas striatal D1 and D2-related genes were predictive of Go/NoGo learning in DEV and IEV conditions, respectively (Frank, Doll, Oas-Terpstra, and Moreno, unpublished observations).

Probability-magnitude bias

Overall, participants were biased to learn more about probability than magnitude of rewards (evidenced by a positive PM-bias). This result is consistent with the existence of a nonlinear utility function described in behavioral economics, whereby higher magnitude gains are preferred to lower gains, but at a declining rate (e.g., Kahneman and Tversky, 1979). Our model embodies a particular set of mechanisms that gives rise to these effects, whereby the basal ganglia enhances the contrast between reinforcement probabilities by subtracting Go and NoGo associations from each other, but underweights large magnitudes (Frank, 2005; Frank and Claus, 2006). Moreover, the diminished PM-bias in medicated patients is accounted for in the model by suppressed learning from negative prediction errors frequent in CEVR (i.e., the same mechanism that leads to reduced performance in IEV).

Conclusion

This study supports the hypothesis that striatal DA effects on temporal decision making and probabilistic selection paradigms tap into common mechanisms, as supported by computational and experimental data. To our knowledge, this is the first study to test the effect of PD and DA medications on this rewarding aspect of temporal decision making.

References

Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9:357–381.

Ashby FG, O'Brien JB (2005) Category learning and multiple memory systems. Trends Cogn Sci 9:83–89.

Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. J Neurophysiol 98:1428–1439.

Berns GS, Sejnowski TJ (1995) How the basal ganglia make decisions. In: Neurobiology of decision-making (Damasio A, Damasio H, Christen Y, eds), pp. 101–113. New York: Springer.

Berridge KC (2007) The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl) 191:391–431.

Brück A, Aalto S, Nurmi E, Vahlberg T, Bergman J, Rinne JO (2006) Striatal subregional 6-[18f]fluoro-l-dopa uptake in early Parkinson's disease: a two-year follow-up study. Mov Disord 21:958–963.

Chevalier G, Deniau JM (1990) Disinhibition as a basic process in the expression of striatal functions. Trends Neurosci 13:277–280.

Cools R (2006) Dopaminergic modulation of cognitive function-implications for L-DOPA treatment in Parkinson's disease. Neurosci Biobehav Rev 30:1–23.

Cools R, Barker RA, Sahakian BJ, Robbins TW (2001) Mechanisms of cognitive set flexibility in Parkinson's disease. Brain 124:2503–2512.

Cools R, Altamirano L, D'Esposito M (2006) Reversal learning in parkinson's disease depends on medication status and outcome valence. Neuropsychologia 44:1663–1673.

Dalley JW, Lääne K, Theobald DE, Armstrong HC, Corlett PR, Chudasama Y, Robbins TW (2005) Time-limited modulation of appetitive pavlovian memory by d1 and nmda receptors in the nucleus accumbens. Proc Natl Acad Sci U S A 102:6189–6194.

Daw ND, Courville AC, Touretzky DS (2003) Timing and partial observability in the dopamine system. In: Advances in neural information processing systems 15: proceedings of the 2002 conference. Cambridge, MA: MIT.

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

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. J Neurophysiol 84:3072–3077.

Delgado MR, Miller MM, Inati S, Phelps EA (2005) An fMRI study of reward-related probability learning. Neuroimage 24:862–873.

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

Egelman DM, Person C, Montague PR (1998) A computational role for dopamine delivery in human decision-making. J Cogn Neurosci 10:623–630.

Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 8:1481–1489.

Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17:51–72.

Frank MJ (2006) Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw 19:1120–1136.

Frank MJ, Claus ED (2006) Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol Rev 113:300–326.

Frank MJ, O'Reilly RC (2006) A mechanistic account of striatal dopamine function in human cognition: Psychopharmacological studies with cabergoline and haloperidol. Behav Neurosci 120:497–517.

Frank MJ, Loughry B, O'Reilly RC (2001) Interactions between the frontal cortex and basal ganglia in working memory: a computational model. Cogn Affect Behav Neurosci 1:137–160.

Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in Parkinsonism. Science 306:1940–1943.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007a) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A 104:16311–16316.

Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007b) Hold your horses: impulsivity, deep brain stimulation and medication in Parkinsonism. Science 318:1309–1312.

Frank MJ, Santamaria A, O'Reilly RC, Willcutt E (2007c) Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. Neuropsychopharmacology 32:1583–1599.

Frank MJ, Scheres A, Sherman SJ (2007d) Understanding decision making deficits in neurological conditions: insights from models of natural action selection. Philos Trans R Soc Biol Sci 362:1641–1654.

Gonon FJ (1997) Prolonged and extrasynaptic excitatory action of dopamine mediated by D1 receptors in the rat striatum *in vivo*. J Neurosci 17:5972–5978.

Graybiel AM (1998) The basal ganglia and chunking of action repertoires. Neurobiol Learn Mem 70:119–136.

Graybiel AM (2004) Network-level neuroplasticity in cortico-basal ganglia pathways. Parkinsonism Relat Disord 10:293–296.

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

Hariri AR, Brown SM, Williamson DE, Flory JD, de Wit H, Manuck SB (2006) Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity. J Neurosci 26:13213–13217.

Heerey EA, Robinson BM, McMahon RP, Gold JM (2007) Delay discounting in schizophrenia. Cognit Neuropsychiatry 12:213–221.

Holroyd CB, Larsen JT, Cohen JD (2004) Context dependence of the event-related brain potential associated with reward and punishment. Psychophysiology 41:245–253.

Houk JC, Bastianen C, Fansler D, Fishbach A, Fraser D, Reber PJ, Roy SA, Simo LS (2007) Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Philos Trans R Soc Lond Biol Sci 362:1573–1583.

Judd CM, McClelland GH (1989) Data analysis, a model-comparison approach. Orlando, FL: Harcourt.

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291.

Kish SJ, Shannak K, Hornykiewicz O (1988) Uneven pattern of dopamine loss in the striatum of patients with idiopathic Parkinson's disease. N Engl J Med 318:876–880.

Knowlton BJ, Mangels JA, Squire LR (1996) A neostriatal habit learning system in humans. Science 273:1399–1402.

Lustig C, Matell MS, Meck WH (2005) Not "just" a coincidence: frontal-striatal interactions in working memory and interval timing. Memory 13: 441–448.

Maddox WT, Filoteo JV (2001) Striatal contributions to category learning: quantitative modeling of simple linear and complex non-linear rule learning in patients with Parkinson's disease. J Int Neuropsychol Soc 7:710–727.

Mazzoni P, Hristova A, Krakauer JW (2007) Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation. J Neurosci 27:7105–7116.

McClure SM, Daw ND, Montague PR (2003) A computational substrate for incentive salience. Trends Neurosci 26:423–428.

McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed rewards. Science 306:503–507.

McClure SM, Ericson KM, Laibson DI, Loewenstein G, Cohen JD (2007) Time discounting for primary rewards. J Neurosci 27:5796–5804.

Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381–425.

Montague PR, Berns GS (2002) Neural economics and the biological substrates of valuation. Neuron 36:265–284.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Moustafa AA, Maida AS (2007) Using TD learning to simulate working memory performance in a model of the prefrontal cortex and basal ganglia. Cogn Syst Res 8:262–281.

Moustafa AA, Sherman SJ, Frank MJ (2008) A dopaminergic basis for working memory, learning, and attentional shifting in Parkinson's disease. Neuropsychologia 46:3144–3156.

Nakamura K, Hikosaka O (2006) Role of dopamine in the primate caudate nucleus in reward modulation of saccades. J Neurosci 26:5360–5369.

Niv Y, Rivlin-Etzion M (2007) Parkinson's disease: fighting the will? J Neurosci 27:11777–11779.

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

O'Reilly RC, Frank MJ, Hazy TE, Watz B (2007) PVLV: the primary value and learned value pavlovian learning algorithm. Behav Neurosci 121:31–49.

Packard MG, McGaugh JL (1996) Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiol Learn Mem 65:65–72.

Pavese N, Evans AH, Tai YF, Hotton G, Brooks DJ, Lees AJ, Piccini P (2006) Clinical correlates of levodopa-induced dopamine release in Parkinson disease: a pet study. Neurology 67:1612–1617.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045.

Reynolds JNJ, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 412:67–69.

Satoh T, Nakai S, Sato T, Kimura M (2003) Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci 23:9913–9923.

Scheres A, Dijkstra M, Ainslie E, Balkan J, Reynolds B, Sonuga-Barke E, Castellanos FX (2006) Temporal and probabilistic discounting of rewards in children and adolescents: effects of age and ADHD symptoms. Neuropsychologia 44:2092–2103.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.

Seeman P (2008) Dopamine d2(high) receptors on intact cells. Synapse 62:314–318.

Shen W, Flajolet M, Greengard P, Surmeier DJ (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321:848–851.

Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA, Poldrack RA (2004) Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. Brain 127:851–859.

Shohamy D, Myers CE, Geghman KD, Sage J, Gluck MA (2006) L-dopa impairs learning, but spares generalization, in Parkinson's disease. Neuropsychologia 44:774–784.

Shohamy D, Myers CE, Kalanithi J, Gluck MA (2008) Basal ganglia and dopamine contributions to probabilistic category learning. Neurosci Biobehav Rev 32:219–236.

Suri RE, Schultz W (1998) Learning of sequential movements by neural network model with dopamine-like reinforcement signal. Exp Brain Res 121:350–354.

Surmeier DJ, 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 Neurosci 30:228–235.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.

Tedroff J, Pedersen M, Aquilonius SM, Hartvig P, Jacobsson G, Långström B (1996) Levodopa-induced changes in synaptic dopamine in patients with Parkinson's disease as measured by [11C]raclopride displacement and PET. Neurology 46:1430–1436.

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. Science 307:1642–1645.

Venton BJ, Zhang H, Garris PA, Phillips PE, Sulzer D, Wightman RM (2003) Real-time decoding of dopamine concentration changes in the caudate-putamen during tonic and phasic firing. J Neurochem 87:1284–1295.

Wickens JR, Begg AG, Arbuthnott GW (1996) Dopamine reverses the depression of rat corticostriatal synapses which normally follows high frequency stimulation of cortex in vitro. Neuroscience 70:1–5.