Behavioral/Cognitive

Impulse Control Disorders in Parkinson's Disease Are Associated with Dysfunction in Stimulus Valuation But Not Action Valuation

Payam Piray,[1] **Yashar Zeighami,**[2] **Fariba Bahrami,**[3] **Abeer M. Eissa,**[4] **Doaa H. Hewedi,**[4] **and Ahmed A. Moustafa**[5]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, 6500 HB Nijmegen, the Netherlands, [2]Montreal Neurological Institute, McGill University, Montreal, QC H3A 2B4, Canada, [3]Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University College of Engineering, University of Tehran, 14395-515 Tehran, Iran, [4]Psychogeriatric Research Center, Department of Psychiatry, School of Medicine, Ain Shams University, 11566 Cairo, Egypt, and [5]School of Social Sciences and Psychology and Marcs Institute for Brain and Behaviour, University of Western Sydney, NSW 2751, Australia

A substantial subset of Parkinson's disease (PD) patients suffers from impulse control disorders (ICDs), which are side effects of dopaminergic medication. Dopamine plays a key role in reinforcement learning processes. One class of reinforcement learning models, known as the actor-critic model, suggests that two components are involved in these reinforcement learning processes: a critic, which estimates values of stimuli and calculates prediction errors, and an actor, which estimates values of potential actions. To understand the information processing mechanism underlying impulsive behavior, we investigated stimulus and action value learning from reward and punishment in four groups of participants: on-medication PD patients with ICD, on-medication PD patients without ICD, off-medication PD patients without ICD, and healthy controls. Analysis of responses suggested that participants used an actor-critic learning strategy and computed prediction errors based on stimulus values rather than action values. Quantitative model fits also revealed that an actor-critic model of the basal ganglia with different learning rates for positive and negative prediction errors best matched the choice data. Moreover, whereas ICDs were associated with model parameters related to stimulus valuation (critic), PD was associated with parameters related to action valuation (actor). Specifically, PD patients with ICD exhibited lower learning from negative prediction errors in the critic, resulting in an underestimation of adverse consequences associated with stimuli. These findings offer a specific neurocomputational account of the nature of compulsive behaviors induced by dopaminergic drugs.

Key words: **computational modeling; dopamine; impulse control disorders; Parkinson's disease; prediction error; reinforcement learning**

Introduction

Dopaminergic medications, especially D2 agonist drugs, trigger impulse control disorders (ICDs) such as hypersexuality, binge eating, and pathological gambling in a subset of Parkinson's disease (PD) patients (Voon et al., 2007). Although PD is primarily associated with dopamine depletion in the substantia nigra and dorsal striatum (Kish et al., 1988), the underlying neural substrates of ICD in PD are mostly the ventral regions of the striatum and their dopaminergic innervations from the ventral tegmental area (Dagher and Robbins, 2009; Voon et al., 2010). Therefore, dopamine neurons projecting to the ventral striatum are relatively intact in PD patients (Kish et al., 1988). Furthermore, it has been suggested that the restoration of dopamine transmission in the dorsal striatum may lead to excessive dopamine receptor stimulation in the ventral striatum (Swainson et al., 2000; Cools et al., 2001), thus inducing ICD in some patients (Cools et al., 2003; Dagher and Robbins, 2009).

Overwhelming evidence has shown that dopamine neurons encode prediction error (PE) signaling, which guides stimulus and action value learning in reinforcement learning (RL) models (Schultz et al., 1997; Bayer and Glimcher, 2005; Pessiglione et al., 2006). It has also been shown that a popular RL model, known as Q-learning (QL), is useful for understanding the mechanistic differences in learning between on- and off-medication PD patients (Frank et al., 2007; Rutledge et al., 2009). Although it has been hypothesized that the functional dissociation of striatal subregions is critical to understanding the underlying mechanism of compulsive behaviors in both the general population (Everitt and Robbins, 2005; Belin et al., 2013) and PD patients (Cools et al., 2007; Dagher and Robbins, 2009), previous RL models of PD have not addressed the different roles of the ventral and dorsal striatum in the development of ICD in PD. A well known RL

Received Sept. 23, 2013; revised April 20, 2014; accepted April 23, 2014.

Author contributions: P.P., Y.Z., F.B., and A.A.M. designed research; A.M.E., D.H.H., and A.A.M. performed research; P.P., Y.Z., and F.B. analyzed data; P.P. and A.A.M. wrote the paper.

This work was partially supported by the University of Western Sydney (2013 Internal Research Grant Scheme Award P00021210 to A.A.M.). We thank Roshan Cools for helpful advice and comments on this manuscript; Ivan Toni and Hanneke den Ouden for helpful conversations and advice; and to Ivan Toni, Sam Gershman, Sean Fallon, Amir Dezfouli, Sylvia Cox, and Stephanie Scala for helpful comments on this manuscript.

Correspondence should be addressed to either of the following: Payam Piray, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands, E-mail: p.piray@donders.ru.nl; or Ahmed A. Moustafa, Marcs Institute for Brain and Behaviour, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia, E-mail: a.moustafa@uws.edu.au.

DOI:10.1523/JNEUROSCI.4063-13.2014

0270-6474/14/347814-11$15.00/0

framework that models the different roles of the dorsal (motor) and ventral (limbic) striatum is the actor-critic (AC) framework (Barto, 1995; Dayan and Balleine, 2002). This framework has two modules, known as the critic and the actor, where the former is responsible for PE computations and stimulus value learning and the latter is responsible for action valuation and selection. Empirical studies suggest that the ventral and the dorsal striatum play different roles in decision making, with the former corresponding to the critic and the latter corresponding to the actor (Cardinal et al., 2002; Packard and Knowlton, 2002; O'Doherty et al., 2004). Based on these neuroanatomical data and a prior AC model of addiction (Piray et al., 2010), we here hypothesize that, whereas PD is associated with the actor (i.e., action valuation and selection), ICDs in PD are associated with the critic (i.e., stimulus valuation and PE computations). Therefore, we provide a novel modeling approach that combines the concept of separate roles for positive and negative PEs in learning (Frank et al., 2007) with the AC framework to test this hypothesis.

Materials and Methods

Participants

This study was part of a larger project conducted at Ain Shams University Hospital, Cairo, Egypt. Participants were asked whether they were willing to participate in the short or long version of the project. In the short version, participants completed 80 trials of a probabilistic learning task compared with 160 trials for the long version. Ninety-five participants were recruited, 79 of which participated in the long version of the project. For this report, we only included those subjects who participated in the long version of the task (with 160 trials). To have the same number of data points across all subjects for estimating the parameters of computational models, we did not include the data from subjects who participated in the short version of the task. This is because, in principle (within-subject) variance of parameters estimated based on 80 trials is larger than those estimated based on 160 trials and this could inflate statistical comparisons between groups.

Data from three participants were discarded from the analysis because these participants had failed to respond in at least 20% of trials. Therefore, 4 groups were included in the analyses: (1) PD patients without ICD tested off medication (PD-OFF, $n = 25$, 6 females); (2) PD patients without ICD tested on medication (PD-ON, $n = 15$, 3 females); (3) PD patients with ICD tested on medication (PD-ON-ICD, $n = 16$, 2 females); and (4) healthy controls ($n = 20$, 7 females). The healthy control participants did not have any history of neurological or psychiatric disorders. All participants gave written informed consent and the study was approved by ethical board of Ain Shams University.

The Unified Parkinson's Disease Rating Scale (UPDRS) was used to measure the severity of PD (Lang and Fahn, 1989). The UPDRS for all patients, including PD-OFF, was measured before the testing session when all PD patients were on medication. There was no difference in UPDRS between the three patient groups ($F_{(2,53)} = 0.29$, $p = 0.75$).

The PD-OFF group was withdrawn from medications for a period of at least 18 h. The majority of on-medication patients were taking dopamine precursors (levodopa-containing medications) and D2 receptor agonists. Specifically, all participants in the PD-ON-ICD group and 14 participants in the PD-ON group were taking D2 agonist medications (either Requip or Mirapex). In addition to D2 agonist medications, 10 patients in the PD-ON-ICD group and 11 patients in the PD-ON group were taking levodopa medications.

All participants were screened for intact cognitive function and absence of dementia with the Mini-Mental Status Exam (MMSE; Folstein et al., 1975). Participants required a score of at least 26 to be considered for the study. All groups were matched for age and education. In addition, we found no difference between the groups on the North American Adult Reading Test (Uttl, 2002), the Beck Depression Inventory (Beck et al., 1987), the MMSE, or the forward and backward digit span tasks (all p-values >0.05, one-way ANOVA). All scales were administered by trained experts (Table 1).

Table 1. Demographic data

	Healthy	PD-OFF	PD-ON	PD-ON-ICD
Age	66.45 (4.70)	63.92 (3.99)	63.33 (3.98)	64.38 (3.32)
Disease duration	NA	9.72 (2.64)	8.87 (3.14)	9.63 (2.45)
HYS	NA	2.54 (0.61)	2.40 (0.57)	2.47 (0.50)
UPDRS	NA	20.36 (5.49)	19.60 (6.42)	19.00 (5.32)
NAART	36.25 (9.15)	34.64 (10.80)	35.60 (12.93)	38.00 (6.32)
MMSE	27.65 (1.18)	27.48 (0.96)	27.00 (0.93)	27.19 (1.11)
Forward DS	6.25 (1.65)	6.80 (1.66)	6.53 (2.13)	6.75 (1.69)
Backward DS	6.25 (1.59)	6.32 (1.80)	6.47 (2.17)	7.00 (1.37)
BDI	7.75 (1.97)	6.92 (1.32)	8.00 (1.69)	6.75 (1.57)
BIS*	54.15 (4.51)	56.80 (4.74)	57.67 (4.18)	61.88 (4.56)

Means are shown with SDs in parentheses.

NA, Not applicable; HYS, Hoehn–Yahr scale; NAART, North American Adult Reading Test; DS, digit span; BDI, Beck Depression Inventory.

*$p < 0.001$.

The diagnosis of ICD was assessed with interviews conducted by neurologists at Ain Shams University Hospital and associated clinics. ICDs reported included compulsive shopping (10 patients), hypersexuality (nine patients), gambling (six patients), and binge eating (four patients). The majority of participants had more than one type of ICD (four patients with only one type of ICD, 11 patients with two ICDs, and one patient with three ICDs). The Barratt Impulsiveness Scale (BIS) was administered to measure trait impulsivity in all groups. There was a highly significant difference in BIS scores between the groups ($F_{(3,72)} = 8.76$, $p < 0.001$). A *post hoc t* test revealed that the effect was mainly driven by a higher impulsivity in the PD-ON-ICD group. BIS scores for this group were significantly higher than those for the other three groups ($p < 0.02$ for all three tests, two-tailed t test). We also found significantly higher BIS scores in the PD-ON group compared with the healthy group ($p < 0.05$, two-tailed t test).

Task

All participants were administered a probabilistic reward and punishment learning task (Fig. 1*A*; Bódi et al., 2009). On each trial, participants viewed one of four different stimuli (S1, S2, S3, and S4) and were asked to decide whether the stimulus belonged to category A or B. Two stimuli (S1 and S2) were used in the reward-learning trials (win or no-win) and the other two stimuli (S3 and S4) were used in the punishment-learning trials (lose or no-lose). Participants received an outcome after making their choices. There was an optimal choice for each stimulus, which predominately resulted in obtaining reward or avoiding punishment (positive feedback; Fig. 1*B*). Therefore, in reward trials, an optimal choice resulted in +25 points 80% of the time and in no reward for 20% of trials. In contrast, a nonoptimal response resulted in +25 points 20% of the time and otherwise resulted in no reward. In punishment trials, an optimal response resulted in −25 points with 20% probability and otherwise resulted in no punishment. In contrast, a nonoptimal response resulted in −25 points 80% of the time and otherwise resulted in no punishment. The task had 160 trials and the order in which stimuli were presented was pseudorandomized in blocks of 40 trials. For every block, each stimulus was randomly presented in 10 trials.

Theoretical framework

We used computational modeling to investigate the mechanistic differences in learning between participant groups. We fitted different RL models to each participant's choice data. These models were variants of either the QL or the AC framework. Notably, QL and AC frameworks use different strategies to calculate the PE, the pivotal signal in learning within both frameworks. Although the QL framework computes the PE signal based on the estimated value of stimulus-action pairs, the AC framework computes the PE based on the estimated value of stimuli regardless of the action taken. The different claims of PE computations in these two frameworks can be examined in a relatively theory-neutral manner through model-independent estimation of PE. We also fitted different models to participants' choices and compared them using Bayesian model comparison. All models use the sequence of choices and feedbacks for every participant to estimate the probability of action taken on every trial.

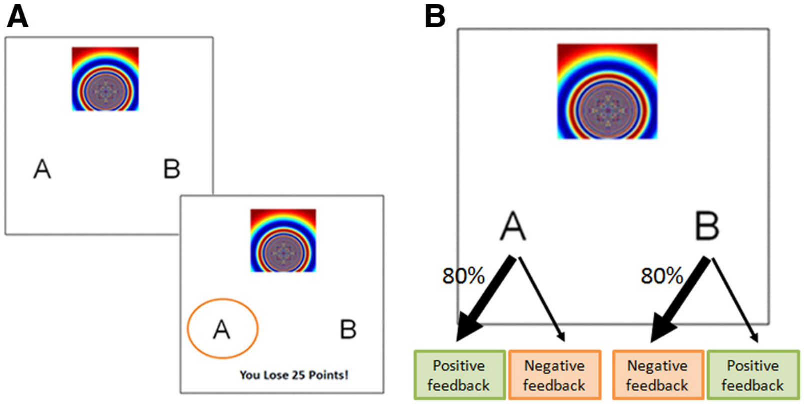

Figure 1. Probabilistic learning task. ***A***, On each trial, one of four stimuli is presented and the participant is asked to choose whether the stimulus belongs to category A or B to avoid punishment or obtain reward. ***B***, Structure of the task. For each stimulus, one of the choices (the optimal action) predominantly (80% of the time) results in a positive feedback (either through obtaining a reward or avoiding punishment). The other choice (the nonoptimal action) predominantly results in a negative feedback.

Reinforcement learning models

The first model is the QL model with different learning rates for positive and negative PEs (dual-α QL; Fig. 2*A*). This model learns the value associated with each stimulus-action pair $Q_t(s_t, a_t)$ using a PE signal, which is the discrepancy between the outcome (reward or punishment) and $Q_t(s_t, a_t)$ as follows:

$$\delta_t = o_t - Q_t(s_t, a_t)$$

where o_t is the outcome on trial t. The model then updates the current estimated value with the PE as follows:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha^+\delta_t \text{ if } \delta_t > 0$$

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha^-\delta_t \text{ if } \delta_t < 0$$

where α^+ and α^- are the learning rates for positive and negative PEs, respectively. These learning rates determine the degree that recent PEs affect the estimated value. If $\alpha^+ > \alpha^-$, then the effect of positive PEs on learned values is larger than that of negative PEs and vice versa if $\alpha^+ < \alpha^-$. The effect of positive and negative PE is equal for $\alpha^+ = \alpha^-$. Frank et al. (2004) hypothesized that different types of dopamine receptors within the striatum mediate the ability to learn from positive and negative PEs via modulation of dopamine activity in the direct and indirect corticostriatothalamic pathways, respectively. According to Frank et al. (2004), the positive PE increases phasic dopamine release, resulting in learning through D1 receptors. Conversely, the negative PE causes a dopamine dip below baseline resulting in learning through D2 receptors (also see Moustafa et al., 2013).

The probability of choosing each action is computed using the soft-max equation:

$$p(c_t = A|s_t) = \frac{1}{1 + \exp[-\beta(Q_t(s_t, A) - Q_t(s_t, B)) - \phi(C_t(s_t, A) - C_t(s_t, B))]}$$

$$p(c_t = B|s_t) = 1 - p(c_t = A|s_t)$$

where $p(c_t = A|s_t)$ and $p(c_t = B|s_t)$ are the probability of choosing A and B, respectively, β is the inverse-temperature parameter that encodes decision noise, and $C_t(s_t, A)$ and $C_t(s_t, B)$ represents the choice of A and B on the last presentation of s_t, respectively (Lau and Glimcher, 2005; Rutledge et al., 2009). Therefore, $C_t(s_t, A) = 1$ and $C_t(s_t, B) = 0$ if A has been chosen in the previous presentation of s_t before trial t, but if B has been chosen, $C_t(s_t, A) = 0$ and $C_t(s_t, B) = 1$. Therefore, ϕ determines the extent to which the previous choice, independent of reward history, affects the current choice. Although positive values of ϕ represent a tendency to perseverate on previous choices, negative values represent a tendency to switch more frequently between available options.

The second model is the AC model (standard AC; Fig. 2*B*), which assigns learning and action selection to two different modules. The PE signal in this model is computed based on stimulus values, regardless of the action taken, as follows:

$$\delta_t = o_t - V_t(s_t)$$

where $V_t(s_t)$ is the current critic's value for s_t. The critic's value is then updated using the PE as follows:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_c\delta_t$$

where α_c is the critic's learning rate. The PE is also conveyed to the actor to update the action value of the selected action in the actor as follows:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_a\delta_t$$

where α_a is the actor's learning rate. Here, if $\alpha_c > \alpha_a$, then the effect of PEs on the critic is larger than that of actor, and vice versa if $\alpha_c < \alpha_a$. Note that this is common practice in machine learning that the update of the actor is slower than that of the critic to ensure that the critic has sufficient time to evaluate the current policy (Grondman et al., 2012). However, we enforce no constraints on the critic's and actor's learning rates. If participants used an AC strategy, we would expect that the fitted parameters satisfy this condition for the majority of participants. The probability of each action is computed according to the actor's action values. A similar soft-max equation as the previous model, dual-α QL, is used to generate the probability of actions based on actor's action values and choice perseveration.

The third model is the dual-α AC model, which is very similar to the standard AC model (Fig. 2*C*). The difference between these two models is how they update stimulus and action values. The dual-α AC model updates stimulus values through two different learning rates, one for positive PEs and one for negative PEs, as follows:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_c^+\delta_t \text{ if } \delta_t > 0$$

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_c^-\delta_t \text{ if } \delta_t < 0$$

If $\alpha_c^+ > \alpha_c^-$, then the effect of positive PEs on the stimulus value is larger than that of negative PEs, and vice versa if $\alpha_c^+ < \alpha_c^-$. The actor's action value is also updated through the two different learning rates for positive and negative PEs as follows:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_a^+\delta_t \text{ if } \delta_t > 0$$

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_a^-\delta_t \text{ if } \delta_t < 0$$

Here, if $\alpha_a^+ > \alpha_a^-$, then the effect of positive PEs on the actor's action value is larger than that of negative PEs, and vice versa if $\alpha_a^+ < \alpha_a^-$. The values for all models were initiated at zero.

Model-independent estimation of PE

In this section, we derive a model-independent estimator of PE. This estimator could then be used to assess learning strategies used by participants in a theory-neutral manner.

RL models often assume that choices are generated using a soft-max equation of action values as follows:

$$p_t(a) = \frac{\exp(\beta Q_t(a))}{\exp(\beta Q_t(a)) + \exp(\beta Q_t(a'))}$$

where a and a' are two available choices and β is the inverse-temperature parameter. $Q_t(a)$ is the action value for a on trial t, which could be generated by either an AC model or by a QL model. Note that Q_t is also a function of state (stimulus) in all of the models. For simplicity (without

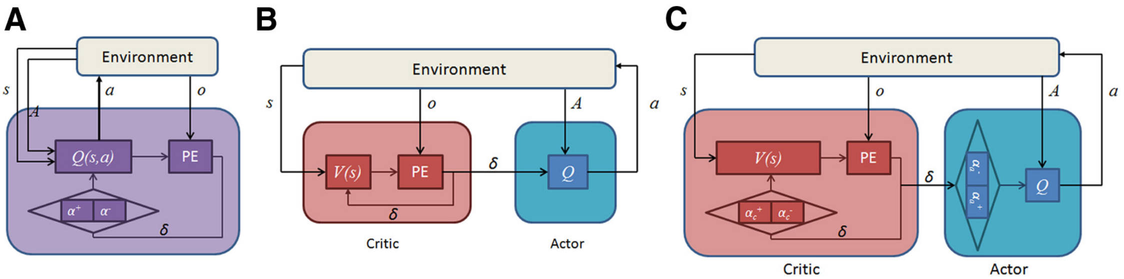

Figure 2. Diagram of the three reinforcement learning models. The environment provides three signals for each model: *s*, indicating the current stimulus, *A*, indicating the set of available actions, and *o*, indicating the outcome after receiving the selected action, *a*, from the model. Every model learns appropriate actions by computing a PE signal (indicated by PE block in the diagram) and selects appropriate actions using estimated Q-values of the available set of actions, *A*. ***A***, The dual-α QL model. This model calculates PEs based on the estimated value of stimulus-selected action pair, *Q*(*s*,*a*). Q-values are updated through two different learning rates, α^+ and α^-, for positive and negative PEs, respectively. ***B***, The standard AC framework: the critic calculates the PE, δ, based on the stimulus value, *V*(*s*), independently from the selected action, *a*. The actor computes action values, *Q*, and selects appropriate action, *a*, from a set of available action, *A*, using actor's Q-values. Both stimulus and action values are updated using the same PE. ***C***, The dual-α AC model. This model has critical features of the previous models. Similar to the standard AC model, the PE is computed based on stimulus values, *V*(*s*), independently from the action, *a*, selected by the actor. Similar to the dual-α QL model, this model updates both the critic's stimulus values, *V*, and the actor's action values, *Q*, through two different learning rates for positive and negative PEs in the critic, α_c^+ and α_c^- and in the actor, α_a^+ and α_a^-.

loss of generality), we focus on sequence of choices related to one state and omit state in the notation in this section. The probability of taking action a' on trial t is computed using a similar equation. Therefore:

$$\frac{p_t(a)}{p_t(a')} = \frac{\exp(\beta Q_t(a))}{\exp(\beta Q_t(a'))}$$

Without loss of generality, we suppose that a is taken at t. Then, the action value of a should be updated using the PE, δ_t, as follows:

$$Q_{t+1}(a) = Q_t(a) + \alpha\delta_t$$

where α is the learning rate. There is no change in the action value of the other action: $Q_{t+1}(a') = Q_t(a')$. Therefore:

$$\frac{p_{t+1}(a)}{p_{t+1}(a')} = \frac{\exp(\beta Q_t(a) + \beta\alpha\delta_t)}{\exp(\beta Q_t(a'))}$$

By subtracting the logarithm of $\frac{p_t(a)}{p_t(a')}$ from the logarithm of $\frac{p_{t+1}(a)}{p_{t+1}(a')}$, we obtain the following:

$$\log\frac{p_{t+1}(a)}{p_{t+1}(a')} - \log\frac{p_t(a)}{p_t(a')} = \beta\alpha\delta_t$$

We define $n_t(a)$ as the number of times that a has been chosen in trials $t' \leq t$. Similarly, $n_t(a)$ is defined as the number of times that a' has been chosen in trials $t' \leq t$. The probability of each choice can be estimated using these variables as follows:

$$p_t(a) \approx \frac{n_t(a)}{n_t(a) + n_t(a')}$$

Accordingly, if $n_t(a') \neq 0$, then $\beta\alpha\delta_t$ can be estimated as follows:

$$\beta\alpha\delta_t \approx \varepsilon_t = \log\frac{n_{t+1}(a)}{n_{t+1}(a')} - \log\frac{n_t(a)}{n_t(a')}$$

where a is the action taken at t and ε_t is the estimator of the PE, which is a quantity that is independent of any specific learning strategy and is purely based on the sequence of choices. Note that the predictions of this estimator match well with the concept of PE. First, if a is chosen in trials t and $t + 1$, then ε_t is positive, suggesting that choosing a resulted in a positive feedback and increased the probability of choosing a for subsequent trials. If a is chosen at t, but not at $t + 1$, then ε_t is negative, suggesting that choosing a resulted in a negative feedback and a reduced the probability of choosing a for future trials. In addition, the magnitude of ε_t is smaller for larger amounts of $n_t(a)$, which is consistent with the idea that the magnitude of PEs should decrease over time.

Subjective utility and nonlearning models

We also fitted four additional models to participants' choices to investigate whether nonlinearity in subjective values of different outcomes, or some nonlearning strategies, could explain data better than the previously mentioned RL models.

Utility models. We considered two utility models. These models test the hypothesis that participants' choices can be explained by nonlinearity in subjective value of outcomes. For the probabilistic learning task used in our study, the subjective value refers to the different subjective utilities for reward and punishment.

The first model is the utility QL model as implemented by Niv et al. (2012). In this model, the PE is computed based on a nonlinear function of the outcomes as follows:

$$\delta_t = U(o_t) - Q_t(s_t, a_t)$$

where $U(o_t)$ is the subjective utility of outcome at time t. The action value is then updated using this PE as follows:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\delta_t$$

Similar to Niv et al. (2012), to model the subjective utility of the outcome, we assumed (without loss of generality) that $U(0) = 0$, $U(-25) = -25$ and $U(+25) = 25u$, where u is a free parameter that determines the subjective utility of outcome. Values of u that are smaller than 1 are consistent with hypersensitivity to punishment, whereas values of u that are larger than 1 are consistent with hypersensitivity to reward. Note that reward and punishment are different from positive and negative PEs that could occur in both reward and punishment trials. This model computes the probability of each action in the same way as the dual-α QL model.

It is also possible to define a subjective utility version of the AC model (utility AC). In this model, the PE is as follows:

$$\delta_t = U(o_t) - V_t(s_t)$$

This PE would then be used by the critic and the actor to update stimulus and action values, respectively. Again, we assumed that $U(0) = 0$, $U(-25) = -25$, and $U(+25) = 25u$, where u is a free parameter that determines the subjective utility of outcome. Similar to the standard AC, two different learning rates are used to update the critic's stimulus values and the actor's action values. This model computes the probability of each action in the same way as the standard AC.

Win-stay lose-shift model. We also considered a model that implemented a win-stay, lose-shift (WSLS) strategy. This model selects actions based only on the most recent outcome. The WSLS strategy selects the same action that led to success on the next trial or chooses a different option on the next trial when an action did not lead to a success. This

Table 2. Bayesian model selection

	No. of free parameters	Healthy	PD-OFF	PD-ON	PD-ON-ICD
Standard AC	3	1653.7	2195.8	1217.5	1344.3
Dual-α QL	4	1660.9	2208.9	1212.8	1299.2
Dual-α AC	5	1587.9	2091.6	1171.4	1289.7
Utility QL	4	1687.5	2182.9	1239.8	1306.7
Utility AC	4	1657.4	2180.1	1212.8	1303.6
WSLS	2	1893.8	2499.9	1414.3	1531.1
WSLS (fixed W)	1	2124.0	2742.6	1587.1	1683.5

These numbers represent the negative log-likelihood of data in the corresponding group given the associated model. The Bayesian model selection takes into account both the goodness of fit and the generalizability of the models. Lower values are associated with better fits. The dual-α AC model fits better than other models for all four groups.

QL, Q-learning; AC, actor-critic.

strategy can be stochastically modeled using a sigmoid function as follows:

$$p(a_t = a|s_t) = \frac{1}{1 + \exp(-\beta w_t(s_t))}$$

where a is the chosen action in the previous presentation of s_t and $\beta > 0$ encodes decision noise. To model the WSLS strategy, we assumed (without loss of generality) that $w_t = -1$ if the previous presentation of s_t was a lose trial and $w_t = W$ if it was a win trial. $W > 0$ is the parameter that determines the weight of win compared with loss. If $W > 1$, then the effect of win on the subsequent choice is larger than that of loss and vice versa if $W < 1$. The effect of win and loss on subsequent choices is symmetric if $W = 1$. For all positive values of W, the probability of choosing the same action as the previous trial is more than the alternative action if the previous trial was a win trial and less than the alternative action if the previous trial was a loss trial. Note that, in our probabilistic learning task, win trials were those that resulted in obtaining a reward in reward trials or avoiding a punishment in punishment trials. We fitted two WSLS models to participants' choices. For the first model we assumed both β and W were free, and in the second one we fixed W at 1. The values for all models were initiated at zero.

Model fitting procedure

We used a hierarchical Bayesian procedure for fitting models to participants' choices as described in Huys et al. (2011a, 2012). All parameters of the models are assumed to be free (see Table 2 for the number of free parameters in each model) except for β in the three AC models (standard AC, dual-α AC, and utility AC), which was fixed at 1. This is because the probabilities of choices for these models are affected by the product of the learning rate parameter of the actor and β and this is the only way that these parameters affect the likelihood function. These two variables are indeed colinear. To show that fixing β at 1 is statistically justified, we also fitted these models with β as a free parameter and used the likelihood ratio test to examine whether these models fit significantly better than the same models with β fixed at 1. For all three models, the fits were not significantly improved by having β as a free parameter ($p > 0.9$ for all groups, likelihood ratio test). Accordingly, the standard AC, dual-α AC, and utility AC models have 3, 5, and 4 free parameters, respectively.

In the hierarchical Bayesian procedure, the parameters of an a priori distribution for individual parameters were estimated using participants' choices through the expectation-maximization algorithm (Dempster et al., 1977). This algorithm is a well known method for finding maximum a posteriori, which alternates between an expectation step and a maximization step. We used Laplace approximation (MacKay, 2003) for the expectation step on each iteration. Assuming a normal distribution for individual parameters, θ^i for the ith participant, this method estimates the mean and the variance of the distributions across the whole group, Θ, which serves as an a priori distribution for finding the maximum a posteriori on the next iteration. For example, for the dual-α AC model, the group parameters are as follows:

$$\Theta = [\mu_{\alpha_c^+}, \nu_{\alpha_c^+}, \mu_{\alpha_c^-}, \nu_{\alpha_c^-}, \mu_{\alpha_a^+}, \nu_{\alpha_a^+}, \mu_{\alpha_a^-}, \nu_{\alpha_a^-}, \mu_\phi, \nu_\phi]'$$

where μ and ν indicate the mean and deviance of the corresponding parameter, respectively. The group mean and variance were estimated separately for each group and were used to define an a priori Gaussian distribution for individual parameters. Therefore, four sets of parameters, associated with four groups, were estimated. For the details of the hierarchical fitting procedure, please refer to Huys et al. (2012).

Bayesian model selection

We used a Bayesian model selection approach to assess which model better captures participants' choices. This approach selects the most parsimonious model by balancing between model fits and different levels of complexity of the models (Kass and Raftery, 1995; MacKay, 2003).

We computed approximate model evidence, $P(D|M)$, which is the probability of participants' choices, D, given the model M. We approximated $P(D|M)$ in log-space using the Bayesian Information Criterion:

$$-\log P(D|M) \approx -\log P(D|M, \Theta_{ML}) + \frac{1}{2}|\Theta|\log|D|$$

where D is the set of all participants' choices in the group, $|D|$ is the number of choices for the whole group and $|\Theta|$ is the number of group parameters. Θ_{ML} is obtained using maximum likelihood as follows:

$$\Theta_{ML} = \arg\max_{\Theta} P(D|M, \Theta)$$

Because Θ_{ML} determines an a priori distribution for individual parameters, we can obtain $P(D|M, \Theta_{\mathrm{ML}})$ using the Laplace approximation as follows:

$$-\log P(D|M, \Theta_{ML}) \approx -\sum_i \log P(D_i|M, \Theta_{ML}, \theta^i_{MAP})$$
$$-\sum_i \log P(\theta^i_{MAP}|\Theta_{ML}) - \frac{1}{2}\sum_i |\theta^i| \log 2\pi + \frac{1}{2}\sum_i \log |H_i|$$

where D_{i} is the set of ith subject's choices, $|\theta^{\mathrm{i}}|$ is the number of free parameters in the model for ith subject, $|H_{\mathrm{i}}|$ is the determinant of the Hessian matrix for ith subject at θ^i_{MAP}, and θ^i_{MAP} is the maximum a posteriori of parameters for the ith subject as follows:

$$\theta^i_{MAP} = \arg\max_{\theta} P(D_i|M, \Theta_{ML}, \theta)\, P(\theta|\Theta_{ML})$$

Model selection using cross-validation

We also performed a cross-validation analysis as a control analysis for model selection. Parameters of the models were fitted based on a subset of choices and generalization of models were assessed by quantifying the prediction probability of the models on a different subset of choices that was not used for fitting (see Daw (2011) for shortcomings of this method in learning studies). Similar to Camerer and Ho (1999), the parameters of models were estimated based on the first two-thirds of trials using the hierarchical Bayesian fitting procedure. Next, the negative log-likelihood of the prediction probability of choices on the remaining one-third of trials was computed and reported.

Statistical analyses

Due to non-Gaussian statistics (because some parameters are expected to lie in the unit range), we used the nonparametric Wilcoxon test for parameter comparison between groups. To ensure that between-group differences were not dependent on parameter regularization used in the hierarchical Bayesian procedure (Wunderlich et al., 2012), we used a permutation test approach as a control analysis. For each significant between-group difference, the labels of the groups were randomly permuted 200 times across the participants of both groups. The parameters for these two pseudorandom groups were then found using the hierarchical Bayesian procedure. We then tested whether the effect size in the real data (assessed by the difference in the median of two groups' parameters) was more than the effect size for the pseudorandom groups.

We also examined between-group differences in stimulus values for both reward and punishment trials. Each subject's fitted parameter values were used to estimate the value of stimuli. The nonparametric Wilcoxon test was used to test between-group differences. A similar control

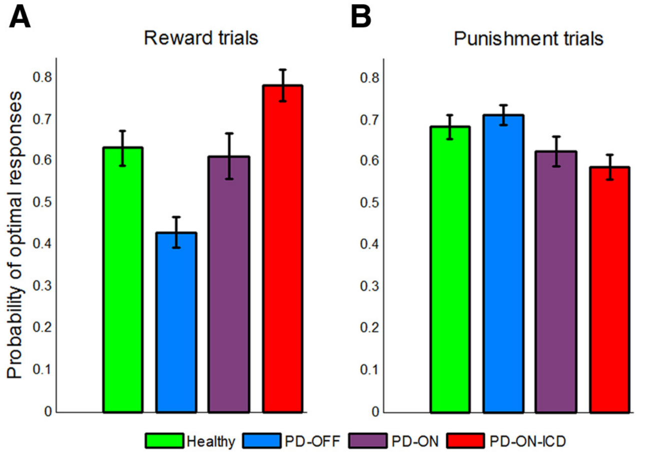

Figure 3. Performance of the four groups on the probabilistic learning task. Shown is mean performance in reward trials (***A***) and punishment trials (***B***). For reward trials, the PD-ON group performed better than the PD-OFF group, but worse than the PD-ON-ICD group. The opposite pattern of performance was observed in punishment trials. Error bars reflect SE.

analysis was also conducted to ensure that the results were not dependent on parameter regularization. Because it is not possible to test between-group differences in stimulus values using the permutation test (due to the dependency of stimulus values in the last presentation of each stimulus on both fitted parameters and sequence of outcomes received), we refitted the dual-α AC model to participants' choices using the hierarchical Bayesian procedure but with only one a priori distribution defined across all participants. Because individual parameters were obtained using the same a priori, the between-group differences cannot be attributed to parameter regularization.

Results

Behavioral data

The probability of optimal responses made by participants was analyzed using an ANOVA with group (four levels: PD-OFF, PD-ON, PD-ON-ICD, and healthy controls) as a between-subject factor and valence (reward or punishment) as a within-subject factor (Fig. 3). This analysis revealed a highly significant interaction between group and valence ($F_{(3.0,72.0)} = 15.81$, $p < 0.001$), as well as a significant main effect of group ($F_{(3.0,72.0)} = 3.79$, $p < 0.05$), but no significant main effects of valence ($F_{(1.0,72.0)} = 2.23$, $p = 0.14$). Further analyses with the additional factor block (two levels: the first half and the second half of the 160 trials) were conducted to assess learning effects. This analysis revealed a significant main effect of block ($F_{(1.0,72.0)} = 14.25$, $p < 0.001$), but no interaction between block and other factors (refer to Fig. 4 for learning curve).

Next, we broke down the significant group by valence interaction into simple main effects of group for the reward and punishment trials separately. All p-values are from two-tailed t test. Therefore, reward learning was impaired in the PD-OFF group relative to the other three groups (healthy controls, PD-ON and PD-ON-ICD groups: $p < 0.001$, $p < 0.01$, and $p < 0.001$, respectively). Conversely, the PD-ON-ICD group showed better reward learning than the other three groups (with healthy controls: $p = 0.015$; with PD-ON: $p = 0.016$).

The opposite pattern of performance was observed for punishment learning. The PD-OFF group exhibited better punishment learning than the PD-ON-ICD ($p = 0.003$) and PD-ON ($p = 0.046$) groups, although there was no significant difference in punishment learning between PD-OFF and healthy participants ($p = 0.43$). Moreover, punishment learning was impaired in the PD-ON-ICD group relative to the healthy control group ($p = 0.028$), although not relative to the PD-ON group ($p = 0.41$).

Model-independent evaluation of learning strategy

Two different strategies could be used to compute the learning signal in the probabilistic learning task. First, the PE could be computed based on the outcome received regardless of which action was taken. This strategy is used by the AC framework. The second strategy is to compute the PE based on the value of the action taken. This strategy is used by the QL framework. The probabilistic learning task allowed us to distinguish between these two learning strategies. For example, if the percentage of optimal responses is 70%, then the critic's stimulus value is affected by the outcomes of both actions and its value (after sufficient trials) is in the middle of two actions' values estimated by the QL framework. For a rewarding stimulus such as S1, the QL value of action A (optimal action), the QL value of action B and AC stimulus values are ~20, 5, and 15.5, respectively. Therefore, if taking an action results in a positive feedback (an outcome of 25 points), then the PE computed by AC is 9.5, but the PE by QL is either 5 or 20 depending on which action is taken. In addition, if taking an action results in a negative feedback (an outcome of 0 points), then the PE computed by the AC is −15.5, but the PE computed by QL is either −20 or −5 depending on the action selected. Therefore, two key events may influence learning signal in this task: whether feedback was positive or negative and whether the action taken was optimal or nonoptimal. Figure 5, *A* and *B*, illustrate the simulated learning signal predicted by the QL and AC frameworks, respectively. As these figures show, whereas both strategies predict a main effect of the feedback, the predictions of the two frameworks are different in terms of the action. Although the AC framework predicts no main effect of action, the QL framework predicts the opposite.

To assess learning strategies used by participants in a relatively theory-neutral manner, we assessed directly the effects of feedback and action on the model-independent estimated PEs across participants (see Materials and Methods), a quantity that is purely based on the sequence of choices for each stimulus. We analyzed the model-independent estimated PEs using an ANOVA with feedback and action as within-subject factors and with group as a between-subject factor. This analysis revealed a highly significant main effect of feedback ($F_{(1.0,70.0)} = 38.5$, $p < 0.001$), consistent with the prediction of both QL and AC frameworks. However, there was no main effect of action ($F_{(1.0,70.0)} = 0.37$, $p = 0.55$), suggesting that the learning strategy used by participants is consistent with the AC learning strategy, but not with that of the QL. As predicted by both learning strategies, no interaction between feedback and action was observed ($F_{(1.0,70.0)} = 1.29$, $p = 0.26$). In addition, no main effect of group and no two- or three-way interactions between group and the other factors were observed ($p > 0.5$), suggesting that all groups used the same learning strategy. Therefore, we plotted the model-independent estimated learning signal across participants in all groups in Figure 5*C*.

We further studied the effects of feedback and action separately for each group using an ANOVA with feedback and action as within-subject factors. Consistent with the previous analysis, there was a main effect of feedback in all four groups (all $p < 0.02$). No main effect of action and no interaction were observed for any of the groups (all $p > 0.16$). Together, these findings suggest that that the learning strategy in all groups is consistent with the predictions by the AC framework.

Note that this analysis holds for the different variants of QL and AC frameworks. Specifically, whereas the dual-α AC model predicts no main effect of action on the learning signal, the dual-α QL model predicts a main effect of action. In addition, both models predict a main effect of feedback and neither predicts an interaction between these factors. Therefore, the results of the analysis of model-independent estimated PEs are consistent with dual-α AC claims about PEs, but not with those of the dual-α QL model.

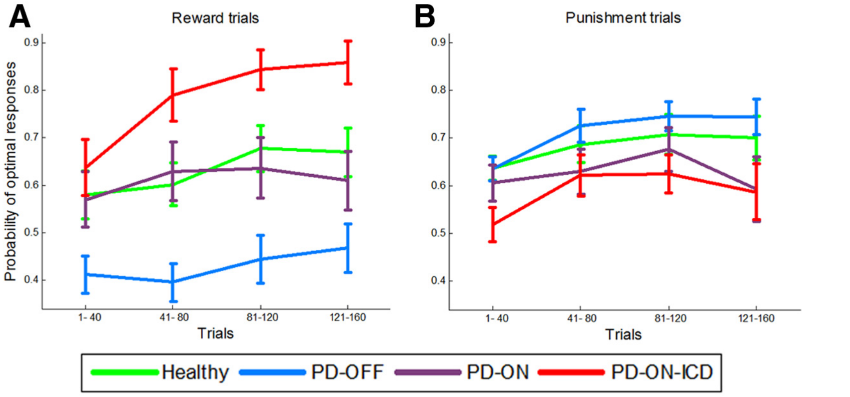

Figure 4. Learning curve for reward trials (***A***) and punishment trials (***B***). The 160 trials are divided in four blocks. Each block contains 20 reward and 20 punishment trials. Error bars indicate SE.

Model comparison

Motivated by these results, we examined the full fit of the models to participants' choices. First, we verified that the models fit significantly better than chance; they did so at $p < 0.001$ for all four groups (likelihood ratio tests). Then, Bayesian model comparison was conducted to identify the best model in each group (Table 2). As Table 2 shows, the negative log-model evidence is lower (with log-Bayes factor of at least 9.5) for the dual-α AC than for the other models for all groups, providing compelling support that the dual-α AC model best captures participants' choices. In the Bayesian model comparison literature, a log-Bayes factor of >3 is taken as strong evidence (cf. the $p < 0.05$ criterion often used in classical statistics; Kass and Raftery, 1995; Daw, 2011). As Table 2 shows, the smallest difference in log-evidence between the best (dual-α AC) and the second best model (dual-α QL) is the one for the PD-ON-ICD group. Because this group is the critical group in this study, we also used a cross-validation approach as a control analysis to compare the plausibility of these two models for this group. Therefore, parameters were fitted based on the first two-thirds of trials and performance of the models quantified on the remaining unseen one-third of trials (Camerer and Ho, 1999). The negative log-likelihood for the dual-α AC and the dual-α QL on the testing dataset were 478.4 and 536.6, respectively. Therefore, the results of cross-validation model selection are consistent with those of the Bayesian model selection, demonstrating strong evidence in favor of the dual-α AC model.

Subsequently, we simulated choices by the best model, the dual-α AC model, using the fitted parameters to verify that the dual-α AC model simulates a similar pattern of between-group differences in optimal responses as observed in the behavioral data (plotted in Fig. 3). These simulated choices were then subject to the same two-tailed *t* test comparisons used in the analyses of between-group differences in behavioral performance. Overall, this simulation analysis replicated similar between-group differences as those observed in the empirical data. The performance of the PD-ON-ICD group in reward trials was significantly better than the other groups ($p < 0.01$). In punishment trials, the PD-OFF group performed significantly better than the PD-ON-ICD group ($p = 0.025$), but not when compared with the other two groups ($p > 0.5$). In addition, consistent with the behavioral results, no difference was found between the PD-ON-ICD and the PD-OFF groups in punishment trials ($p = 0.24$). The simulated choices failed to replicate the findings of significant lower performance by PD-OFF compared with healthy controls and PD-ON in reward trials ($p > 0.05$), although the mean performance of PD-OFF was lower than these groups in reward trials.

Between-group differences in the critic and actor

Next, we assessed between-group differences in parameter values of the best model, dual-α AC. Figure 6 shows the learning rates in the critic and the actor. As this figure shows, the actor's learning rates are generally lower than the critic's learning rates. This learning rate profile ensures that the critic has sufficient time to evaluate the current policy exploited by the actor (Grondman et al., 2012).

First, we studied between-group differences in the critic's parameters. According to our hypothesis, we expected an association between ICD and the critic's learning rates. Although there was no significant difference in α_c^+ between PD-ON-ICD and other groups ($p > 0.1$ for all three tests; Fig. 6*A*), we found a significantly lower learning rate from negative PEs in PD-ON-ICD. Indeed, as Figure 6*B* shows, α_c^- in PD-ON-ICD was less than healthy participants ($p = 0.002$), PD-OFF ($p < 0.001$) and PD-ON ($p = 0.017$). No other group differences in α_c^- were found.

We also investigated between-group differences in the actor's learning rates. Based on the previous data (Frank et al., 2004) and our hypothesis that PD is associated with action valuation deficits, we expected a relatively lower learning rate for the positive PE in PD-OFF and a relatively lower learning rate for the negative PE in PD-ON. As Figure 6*C* shows, α_a^+ was significantly lower in PD-OFF than PD-ON ($p = 0.050$). Conversely, α_a^- was higher in PD-OFF than PD-ON, despite showing only a trend toward significance ($p = 0.058$; Fig. 6*D*). Consistent with our hypothesis, there was no significant difference between PD-ON-ICD and PD-ON in terms of the actor's parameters (no difference between PD-ON and PD-ON-ICD for either α_a^+ ($p = 0.35$) or α_a^- ($p = 0.77$)).

Using the AC framework, it is possible to also evaluate stimulus values. Therefore, we derived the value of every stimulus at the end of the task (the last presentation of the stimulus) for each subject using the subject's choices and the fitted parameters in the dual-α AC model (Fig. 7). We then tested between-group differences in stimulus value separately in reward and punishment trials. Note that two stimuli were only presented in reward trials and two other stimuli were only presented in punishment trials. The stimulus value in punishment trials for the PD-ON-ICD group was significantly less negative than those for the PD-OFF ($p = 0.003$), PD-ON ($p = 0.038$), and healthy control ($p = 0.02$) groups, suggesting that PD patients with ICD underestimate the adverse consequences of stimuli associated with punishment. No significant difference in the stimulus value in reward trials be-

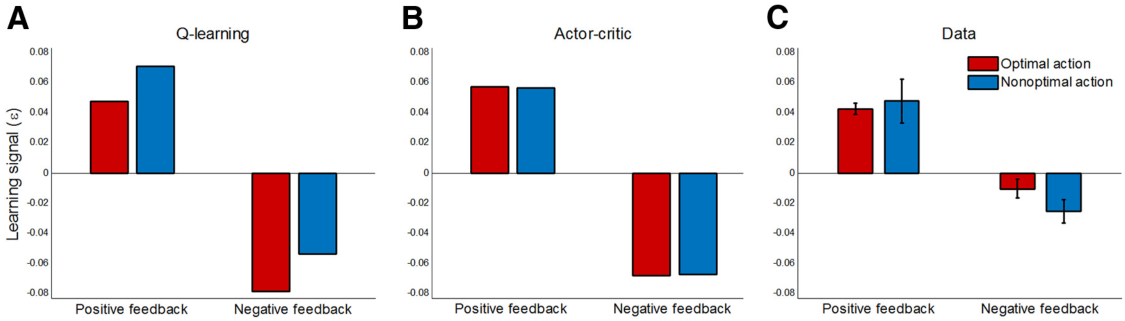

Figure 5. Factorial analysis of model-independent estimates of the learning signal. ***A***, QL framework computes the learning signal based on action values and predicts that this signal depends on whether optimal action or nonoptimal action is taken. ***B***, AC framework computes the learning signal based on the stimulus value regardless of which action is taken. ***C***, Model-independent estimated learning signal based on the data, averaged across participants, is consistent with the prediction of the AC framework. Both models were simulated with learning rates, α, of 0.05 and β inverse temperature of 0.1. The learning signal for both models, ε_t, was defined as $\beta\alpha\delta_t$, where δ_t is the PE computed by the model at trial t. See Materials and Methods for the definition of the model-independent estimates of learning signal. Error bars indicate SE.

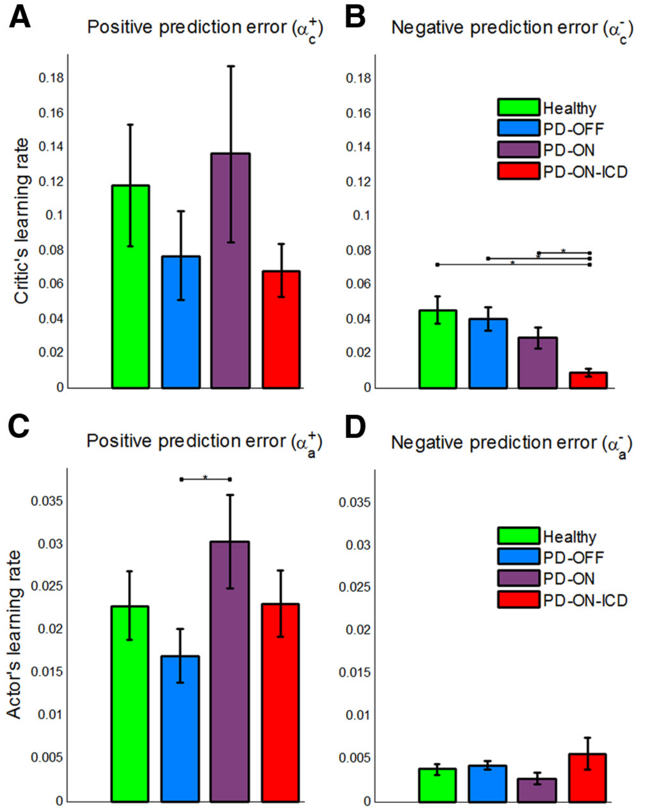

Figure 6. Learning rates in the best model, dual-α AC model. Shown are the critic's learning rate for the positive PE, α_c^+ (***A***), and the negative PE, α_c^- (***B***). PD-ON-ICD showed lower critic's learning rate for the negative PE compared with other three groups, including PD-ON patients. ***C***, ***D***, The actor's learning rate for the positive, α_a^+ (***C***) and the negative PE, α_a^- (***D***). *Significant difference, $p < 0.05$. Error bars indicate SE.

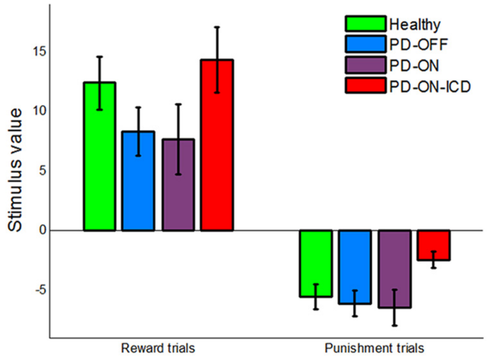

Figure 7. Stimulus value in reward and punishment trials. The stimulus values were obtained using the fitted parameters in the dual-α AC model for the last presentation of each stimulus and averaged across participants. PD-ON-ICD patients exhibited significantly less negative stimulus value in punishment trials compared with the other groups. Error bars indicate SE.

tween PD-ON-ICD and PD-ON was found ($p = 0.35$). Consistent with our hypothesis, the two groups of PD patients without ICD showed a similar pattern of stimulus values in both reward and punishment trials (no difference between PD-OFF and PD-ON for either reward trials, $p = 0.74$, or punishment trials, $p = 0.60$), which supports the idea that PD is not associated with stimulus valuation deficits.

We should note that our main results are independent of the parameter regularization: the learning rate for the negative PE, α_a^-, was significantly lower in the PD-ON-ICD group than in other three groups even when using the permutation test ($p <$ 0.05, two-tailed test). The control analysis for between-group differences in stimulus values also revealed the same significant between-group differences as in our original analysis.

Although the dual-α AC model outperformed the dual-α QL model in all four groups, we also present the results of the between-group difference tests in learning rates for the positive and negative PEs in the dual-α QL model to highlight the benefits of AC modeling for ICD. The learning rate for positive PEs, α^+, was significantly higher in PD-ON compared with healthy controls ($p = 0.002$) and marginally higher in PD-ON compared with PD-OFF ($p = 0.07$). This parameter was also significantly higher in PD-ON-ICD compared with healthy controls ($p < 0.001$). However, there was no difference in α^+ between PD-ON-ICD and PD-ON ($p = 0.51$). There was also no difference between PD-ON-ICD and PD-OFF ($p = 0.15$). No significant between-group differences found in the learning rate for negative PEs, α^- (all $p > 0.6$). Therefore, as these analyses revealed, no difference was found in parameter values of the dual-α QL between on-medication patients with ICD and those without ICD.

Between-group differences in the perseveration

A recent RL study of PD patients reported that the perseveration parameter is dopamine dependent. Therefore, off-medication PD patients exhibited higher perseveration than on-medication patients (Rutledge et al., 2009). Although it is not the main focus of this study, we also examined the effects of the perseveration

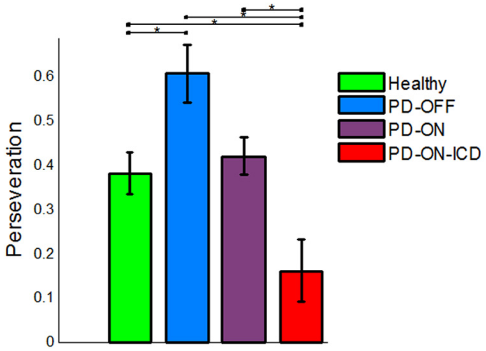

Figure 8. Perseveration parameter. This parameter determines the effect of perseveration on choice. The perseveration parameter depended on dopaminergic medications. In addition, PD-ON-ICD patients exhibited lower perseveration than the other three groups. *Significant difference, $p < 0.05$. Error bars indicate SE.

parameter on model fits and between-group differences in ϕ, the parameter determining the degree that perseveration affects choice. To show that including ϕ in the dual-α AC model is statistically justified, we first tested whether the model with the perseveration parameter fitted significantly better than the same model without the perseveration parameter; it did so for all groups ($p < 0.001$, likelihood ratio test). Note that the perseveration parameter encodes the probability of repeating an action on the subsequent presentation of a stimulus. An alternative way to define perseveration could be to compute the probability of repeating an action on the subsequent trial regardless of the stimulus presented. Therefore, we also fitted a dual-α AC model with the stimulus-independent perseveration and used model selection to test whether the original model outperforms this model; it did so for all groups (with log-Bayes factor of > 6.3).

As Figure 8 shows, consistent with Rutledge et al. (2009), we found significantly higher perseveration values in the PD-OFF group compared with healthy controls ($p = 0.03$). Interestingly, we also found significantly lower perseveration in the PD-ON-ICD group than in the PD-OFF, PD-ON, and healthy control groups ($p < 0.01$ for all three tests). No significant difference was found between the PD-OFF and the PD-ON groups ($p = 0.08$).

Discussion

Dopaminergic medications trigger ICD in a subset of PD patients. In this study, we used a reward and punishment probabilistic learning task and fitted RL models to participants' choices to investigate the mechanistic differences in stimulus valuation and action selection in PD patients with and without ICD. The probabilistic learning task allowed us to distinguish between different learning strategies used by QL and AC frameworks through their different claims about the effects of actions taken on learning. We found that model-independent estimates of the learning signal are consistent with the hallmark of the AC learning strategy. The full fit of models and Bayesian model comparison revealed that an AC model (with different learning rates for positive and negative PEs in both the critic and the actor) best matches participants' choices.

We found that PD patients with ICD (on medication) are more sensitive to rewarding outcomes. Computational modeling revealed that these patients also underestimate adverse consequences of stimuli associated with punishment. We also found computational evidence that patients with ICD exhibit reduced ability in updating stimulus values by negative PEs. Therefore, our findings suggest that distorted stimulus valuation could result in aberrant PE signals, which subsequently affects action values.

There is a great deal of evidence that the ventral striatum contributes to decision making in a manner consistent with the role of the critic in stimulus valuation and PE computations (Cardinal et al., 2002; Dayan and Balleine, 2002; Packard and Knowlton, 2002; O'Doherty et al., 2004). Therefore, our findings are consistent with previous studies that found dopamine-dependent ventral striatal dysfunction in PD patients with ICD symptoms (Cools et al., 2007; Dagher and Robbins, 2009; Steeves et al., 2009; Voon et al., 2010). For example, in a [^{11}C] raclopride positron emission tomography study of PD patients with and without pathological gambling, Steeves et al. (2009) found greater decreases in binding potential in the ventral striatum in on medication PD patients with pathological gambling. In addition, Voon et al. (2010) reported impaired PE signaling in the ventral striatum of PD patients with ICD.

We also found that PD (without ICD) is associated with parameters related to action valuation, but not with stimulus valuation. Therefore, although PD patients without ICD exhibited no deficit in learning stimulus value used for calculating PEs, they showed abnormalities in updating action values with the information signaled by the critic. Therefore, our findings suggest that PD patients without ICD have relatively intact PE computations (in their relatively intact ventral striatum), but the effects of PEs on action values are distorted (in their severely depleted dorsal striatum). These findings are consistent with the hypothesis that the dorsal striatum, the most affected striatal region in PD, is responsible for action valuation and selection. In addition, we also found that the action valuation abnormalities in PD patients without ICD interact with dopaminergic medications. Therefore, consistent with previous data (Frank et al., 2004; Moustafa et al., 2008, 2013; Bódi et al., 2009), we found that whereas off-medication PD patients were better at learning from punishment, on-medication PD patients were better at learning from reward. Mechanistically, we found that off-medication patients, compared with on-medication patients, showed lower action value learning from positive PEs and marginally higher action value learning from negative PEs. Notably, almost all patients in this study received D2 agonist medications, which stimulate D2 dopamine receptors. Therefore, this finding is consistent with the hypothesis of Frank et al. (2004) that different types of dopamine receptors within the striatum, especially those in more dorsal regions, mediate the ability to learn from positive and negative PEs via modulation of dopamine activity in the direct and indirect basal ganglia pathways, respectively (Frank et al., 2004, 2007; O'Reilly et al., 2007). According to this hypothesis, the positive PE increases phasic dopamine release, which facilitates learning by acting on D1 receptors. Conversely, the negative PE results in a dopamine dip below baseline, which facilitates learning by acting on D2 receptors.

Although the role of D1 and D2 receptors in the ventral striatal region, especially the nucleus accumbens shell, is less clear than in the dorsal striatum (Ikemoto et al., 1997; Hopf et al., 2003), there is increasing evidence that the ventral striatal D2 receptors are also involved in learning from negative PEs. Indeed, the negative PE results in dopamine dips below baseline (Bayer and Glimcher, 2005; Hart et al., 2014), which can stimulate high-affinity D2 receptors, but not D1 receptors (Frank et al., 2004). It has also been suggested that D2, but not D1, receptors are stimulated with tonic dopamine release (Grace, 1991). Therefore, as noted by Frank et al. (2004), D2 agonist drugs might fill the dips and reduce the ability to learn from negative PEs. In rats, nucleus accumbens D2 stimulation with a dopamine agonist reduced the ability to learn from negative feedback (Goto and Grace, 2005). In addition, Al carriers of the TAQ-1A polymorphism, which is associated with a lower density of striatal D2 receptors, showed impaired learning from negative feedbacks and aberrant reward-related responses in the ventral striatum (Klein et al., 2007). This hypothesis is consistent with data reporting that ICDs are observed more often in patients on D2 agonist medications (Weintraub et al., 2006; Voon et al., 2007).

An important open question is which individual differences in PD patients with ICD interact with D2 agonist medications and induce compulsive behaviors. One possible answer is that patients vulnerable to ICD have a lower ventral striatal D2 receptor density even before the onset of PD (Dagher and Robbins, 2009). There is limited but important evidence from animal models of cocaine addiction that rats with lower nucleus accumbens D2 receptor density are more impulsive, even before cocaine exposure (Dalley et al., 2007), and are more likely to develop compulsive drug seeking (Belin et al., 2008). In addition, Weintraub et al. (2006) investigated ICD in a large sample of PD patients and reported that those with ICDs were more likely to have had ICDs before the onset of PD. Moreover, animal model studies of addiction have reported that drug exposure further reduces striatal D2 receptors (Nader et al., 2002; Porrino et al., 2004). Similarly, the overstimulation of the ventral striatum in PD patients by D2 agonist medications may further reduce the density of ventral striatal D2, making them more susceptible to develop ICD. Consistent with these ideas, it has been reported that PD patients with ICD showed lower density of D2 receptors in the ventral striatum (Steeves et al., 2009), although it is not clear from this particular study that the reduced level of D2 receptors in the ventral striatum is a predisposing neurobiological trait and/or a consequence of medication.

In summary, we found that whereas PD is associated with parameters related to action valuation and selection, ICDs in PD are mechanistically associated with parameters related to stimulus valuation and PE computations. Specifically, we found computational evidence that ICDs in PD are associated with lower learning rates from negative feedbacks in the critic. These findings offer a computational interpretation of ICDs in PD and highlight the value of computational modeling in understanding cognitive deficits associated with psychiatric disorders (Redish et al., 2008; Huys et al., 2011b; Maia and Frank, 2011; Montague et al., 2012; Monterosso et al., 2012).

References

Barto AG (1995) Adaptive critic and the basal ganglia. In: Models of information processing in the basal ganglia (Houk JC, Davis JL, Beiser DG, eds), pp 215–232. Cambridge: MIT.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141. CrossRef Medline

Beck DC, Carlson GA, Russell AT, Brownfield FE (1987) Use of depression rating instruments in developmentally and educationally delayed adolescents. J Am Acad Child Adolesc Psychiatry 26:97–100. CrossRef Medline

Belin D, Mar AC, Dalley JW, Robbins TW, Everitt BJ (2008) High impulsivity predicts the switch to compulsive cocaine-taking. Science 320:1352–1355. CrossRef Medline

Belin D, Belin-Rauscent A, Murray JE, Everitt BJ (2013) Addiction: failure of control over maladaptive incentive habits. Curr Opin Neurobiol 23:564–572. CrossRef Medline

Bódi N, Kéri S, Nagy H, Moustafa A, Myers CE, Daw N, Dibó G, Takáts A, Bereczki D, Gluck MA (2009) Reward-learning and the novelty-seeking personality: a between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. Brain 132:2385–2395. CrossRef Medline

Camerer C, Ho T-H (1999) Experience-weighted attraction learning in games: a unifying approach. Econometrica 67:827–874. CrossRef

Cardinal RN, Parkinson JA, Hall J, Everitt BJ (2002) Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. Neurosci Biobehav Rev 26:321–352. CrossRef Medline

Cools R, Barker RA, Sahakian BJ, Robbins TW (2001) Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. Cereb Cortex 11:1136–1143. CrossRef Medline

Cools R, Barker RA, Sahakian BJ, Robbins TW (2003) L-Dopa medication remediates cognitive inflexibility, but increases impulsivity in patients with Parkinson's disease. Neuropsychologia 41:1431–1441. CrossRef Medline

Cools R, Lewis SJ, Clark L, Barker RA, Robbins TW (2007) L-DOPA disrupts activity in the nucleus accumbens during reversal learning in Parkinson's disease. Neuropsychopharmacology 32:180–189. CrossRef Medline

Dagher A, Robbins TW (2009) Personality, addiction, dopamine: insights from Parkinson's disease. Neuron 61:502–510. CrossRef Medline

Dalley JW, Fryer TD, Brichard L, Robinson ES, Theobald DE, Lääne K, Peña Y, Murphy ER, Shah Y, Probst K, Abakumova I, Aigbirhio FI, Richards HK, Hong Y, Baron JC, Everitt BJ, Robbins TW (2007) Nucleus accumbens D2/3 receptors predict trait impulsivity and cocaine reinforcement. Science 315:1267–1270. CrossRef Medline

Daw ND (2011) Trial-by-trial data analysis using computational models. In: Decision making, affect, and learning: attention and performance XXIII (Delgado MR, Phelps EA, Robbins TW, eds), pp 3–38. New York: OUP.

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298. CrossRef Medline

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39:1–38.

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

Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12:189–198. CrossRef Medline

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

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

Goto Y, Grace AA (2005) Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. Nat Neurosci 8:805–812. CrossRef Medline

Grace AA (1991) Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience 41:1–24. CrossRef Medline

Grondman I, Busoniu L, Lopes GAD, Babuska R (2012) A survey of actor-critic reinforcement learning: standard and natural policy gradients. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42:1291–1307.

Hart AS, Rutledge RB, Glimcher PW, Phillips PE (2014) Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci 34:698–704. CrossRef Medline

Hopf FW, Cascini MG, Gordon AS, Diamond I, Bonci A (2003) Cooperative activation of dopamine D1 and D2 receptors increases spike firing of nucleus accumbens neurons via G-protein betagamma subunits. J Neurosci 23:5079–5087. Medline

Huys QJ, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, Dayan P (2011a) Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. PLoS Comput Biol 7:e1002028. CrossRef Medline

Huys QJ, Moutoussis M, Williams J (2011b) Are computational models of any use to psychiatry? Neural Netw 24:544–551. CrossRef Medline

Huys QJ, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP (2012) Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol 8:e1002410. CrossRef Medline

Ikemoto S, Glazier BS, Murphy JM, McBride WJ (1997) Role of dopamine D1 and D2 receptors in the nucleus accumbens in mediating reward. J Neurosci 17:8580–8587. Medline

Kass RE, Raftery AE (1995) Bayes factor. Journal of the American Statistical Association 90:773–795. CrossRef

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

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M (2007) Genetically determined differences in learning from errors. Science 318:1642–1645. CrossRef Medline

Lang AE, Fahn S (1989) Assesment of Parkinson's disease. In: Quantification of neurologic deficit (Munsat TL, ed), pp 285–309. Boston: Butterworths.

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579. CrossRef Medline

MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge: Cambridge University.

Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154–162. CrossRef Medline

Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16:72–80. CrossRef Medline

Monterosso J, Piray P, Luo S (2012) Neuroeconomics and the study of addiction. Biol Psychiatry 72:107–112. CrossRef Medline

Moustafa AA, Cohen MX, Sherman SJ, Frank MJ (2008) A role for dopamine in temporal decision making and reward maximization in parkinsonism. J Neurosci 28:12294–12304. CrossRef Medline

Moustafa AA, Krishna R, Eissa AM, Hewedi DH (2013) Factors underlying probabilistic and deterministic stimulus-response learning performance in medicated and unmedicated patients with Parkinson's disease. Neuropsychology 27:498–510. CrossRef Medline

Nader MA, Daunais JB, Moore T, Nader SH, Moore RJ, Smith HR, Friedman DP, Porrino LJ (2002) Effects of cocaine self-administration on striatal dopamine systems in rhesus monkeys: initial and chronic exposure. Neuropsychopharmacology 27:35–46. CrossRef Medline

Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci 32:551–562. CrossRef Medline

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454. CrossRef Medline

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. CrossRef Medline

Packard MG, Knowlton BJ (2002) Learning and memory functions of the Basal Ganglia. Annu Rev Neurosci 25:563–593. CrossRef Medline

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. CrossRef Medline

Piray P, Keramati MM, Dezfouli A, Lucas C, Mokri A (2010) Individual differences in nucleus accumbens dopamine receptors predict development of addiction-like behavior: a computational approach. Neural Comput 22:2334–2368. CrossRef Medline

Porrino LJ, Daunais JB, Smith HR, Nader MA (2004) The expanding effects of cocaine: studies in a nonhuman primate model of cocaine self-administration. Neurosci Biobehav Rev 27:813–820. CrossRef Medline

Redish AD, Jensen S, Johnson A. A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31:415–437, 2008; discussion 437–487.

Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA, Glimcher PW (2009) Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J Neurosci 29:15104–15114. CrossRef Medline

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

Steeves TD, Miyasaki J, Zurowski M, Lang AE, Pellecchia G, Van Eimeren T, Rusjan P, Houle S, Strafella AP (2009) Increased striatal dopamine release in parkinsonian patients with pathological gambling: a [11C] raclopride PET study. Brain 132:1376–1385. CrossRef Medline

Swainson R, Rogers RD, Sahakian BJ, Summers BA, Polkey CE, Robbins TW (2000) Probabilistic learning and reversal deficits in patients with Parkinson's disease or frontal or temporal lobe lesions: possible adverse effects of dopaminergic medication. Neuropsychologia 38:596–612. CrossRef Medline

Uttl B (2002) North American Adult Reading Test: age norms, reliability, and validity. J Clin Exp Neuropsychol 24:1123–1137. CrossRef Medline

Voon V, Potenza MN, Thomsen T (2007) Medication-related impulse control and repetitive behaviors in Parkinson's disease. Curr Opin Neurol 20:484–492. CrossRef Medline

Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, Hallett M (2010) Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. Neuron 65:135–142. CrossRef Medline

Weintraub D, Siderowf AD, Potenza MN, Goveas J, Morales KH, Duda JE, Moberg PJ, Stern MB (2006) Dopamine Agonist Use is Associated with Impulse Control Disorders in Parkinson's Disease. Arch Neurol 63:969–973. CrossRef Medline

Wunderlich K, Smittenaar P, Dolan RJ (2012) Dopamine enhances model-based over model-free choice behavior. Neuron 75:418–424. CrossRef Medline