Neuron
Report

Dopamine Enhances Model-Based over Model-Free Choice Behavior

Klaus Wunderlich,[1,2,*] Peter Smittenaar,[1,2] and Raymond J. Dolan[1]
[1]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK
[2]These authors contributed equally to this work
*Correspondence: kwunder@gmail.com
http://dx.doi.org/10.1016/j.neuron.2012.03.042

SUMMARY

Decision making is often considered to arise out of contributions from a model-free habitual system and a model-based goal-directed system. Here, we investigated the effect of a dopamine manipulation on the degree to which either system contributes to instrumental behavior in a two-stage Markov decision task, which has been shown to discriminate model-free from model-based control. We found increased dopamine levels promote model-based over model-free choice.

INTRODUCTION

An overarching view of adaptive behavior is that humans and animals act to maximize reward and minimize punishment as a consequence of their choices. There are multiple ways this can be realized, and mounting evidence indicates model-based and model-free forms of reinforcement learning (RL) contribute to behavioral control (Balleine and O'Doherty, 2010; Boureau and Dayan, 2011; Daw et al., 2005; Doya, 1999; Redgrave et al., 2010; Wunderlich et al., 2012). Model-free RL learns the course of action leading to maximum long-run reward through a temporal difference (TD) prediction error teaching signal (Montague et al., 1996). By comparison, model-based choice involves forward planning, in which an agent searches a cognitive model of the environment to find the same optimal actions (Dickinson and Balleine, 2002).

An unresolved question is whether neuromodulatory systems implicated in value-based decision making, specifically dopamine, impact on the degree to which one or the other controller is dominant in choice behavior. Phasic firing of dopaminergic VTA neurons encodes reward prediction errors in reinforcement learning (Hollerman and Schultz, 1998; Schultz et al., 1997). In humans, drugs enhancing dopaminergic function (e.g., L-DOPA) augment a striatal signal that expresses reward prediction errors during instrumental learning and, in so doing, increases the likelihood of choosing stimuli associated with greater monetary gains (Bódi et al., 2009; Frank et al., 2004; Pessiglione et al., 2006).

While previous research has focused on the role of dopamine in model-free learning, and value updating via reward prediction errors, its role in model-based choice remains poorly understood. For example, it is unknown if and how dopamine impacts on performance in model-based decisions and on the arbitration between model-based and model-free controllers. This is the question we address in the present study, in which we formally test whether dopamine influences the degree to which behavior is governed by either control system.

RESULTS

We studied 18 subjects on a two-stage Markov decision task after being treated with Madopar (150 mg L-DOPA plus 37.5 mg benserazide) or a placebo in a double-blind, fully counterbalanced, repeated-measures design. We used a task previously shown to distinguish model-based and model-free components of human behavior and in which subjects' choices pertain to a mixture of both systems (Daw et al., 2011). These properties render this task optimally suited to test the influence of a pharmacological manipulation on the degree to which choice performance expresses model-based or model-free control.

In each trial, subjects made an initial choice between two fractal stimuli, leading to either of two second-stage states in which they made another choice between two different stimuli (see Figures 1A and 1B). Each of the four second-stage stimuli was associated with probabilistic monetary reward. To incentivize subjects to continue learning throughout the task, we changed these probabilities slowly and independently according to Gaussian random walks. The choice of each stimulus on the first stage led predominantly (70% of the time) to one of the two associated second-stage states, a relationship that was fixed throughout the experiment. The logic of the task was that a dependence on model-based or model-free strategies predicts different patterns by which feedback obtained after the second stage should impact future first-stage choices.

We first considered stay-switch behavior as a minimally constrained approach to dissociate model-based and model-free control. A model-free reinforcement learning strategy predicts a main effect of reward on stay probability. This is because model-free choice works without considering structure in the environment; hence, rewarded choices are more likely to be repeated, regardless of whether that reward followed a common or rare transition. A reward after an uncommon transition would therefore adversely increase the value of the chosen first-stage cue without updating the value of the unchosen cue. In contrast,

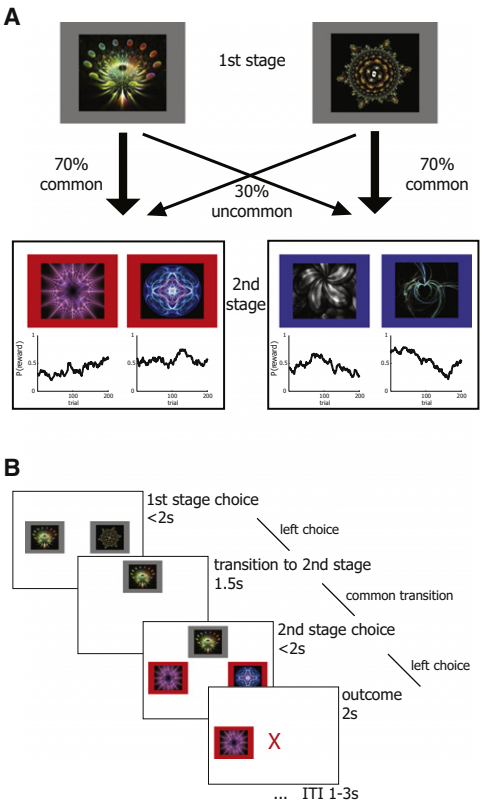

Figure 1. Task Design

Task. (A) On every trial, a choice between two stimuli (left-right randomized) led probabilistically to one of two second-stage states, each of which then demanded another choice between two different stimulus pairs. Importantly, each first-stage stimulus was more strongly (70% versus 30%) associated with a particular second-stage state throughout the experiment, imposing a task structure that could be exploited in model-based choice. All stimuli in stage 2 were associated with probabilistic reward, which changed slowly and independently according to Gaussian random walks. This forced subjects to continuously learn and explore the second-stage choices throughout the experiment. (B) Timings in a single trial. (C) Model-based and model-free strategies for RL predict different patterns by which outcomes experienced after the second stage should impact first-stage choices on subsequent trials (based on Daw et al., 2011). If choices were driven by the model-free system, then a reward should increase the likelihood of choosing the same stimulus on the next trial, regardless of the type of transition (left). Alternatively, if choices were driven by a model-based system, we would expect an interaction between transition type and reward (right).

under a model-based strategy, we expect a crossover interaction between the two factors, because a rare transition inverts the effect of a subsequent reward (Figure 1C). Under model-based control, receiving a reward after an uncommon transition increases the propensity to switch. This is because the rewarded second-stage stimulus can be more reliably accessed by choosing the rejected first-stage cue than by choosing the same cue again.

Using repeated-measures ANOVA, we examined the probability of staying or switching at the first stage dependent on drug state (L-DOPA or placebo), reward on previous trial (reward or no reward), and transition type on previous trial (common or uncommon) (see Figure 2A). A significant main effect of reward, $F(1,17) = 23.3$, $p < 0.001$, demonstrates a model-free component in behavior (i.e., reward increases stay probability regardless of the transition type). A significant interaction between reward and transition, $F(1,17) = 9.75$, $p = 0.006$, reveals a model-based component (i.e., subjects also take the task structure into account). These results show both a direct reinforcement effect (model-free) and an effect of task structure (model-based) and replicate previous findings (Daw et al., 2011).

The key analyses here concerned whether L-DOPA modulated choice propensities. Critically, we observed a significant drug × reward × transition interaction, $F(1,17) = 9.86$, $p = 0.006$, reflecting increased model-based behavior under L-DOPA treatment. We also observed a main effect of the drug, $F(1,17) = 7.04$, $p = 0.017$, showing that subjects are less perseverative under L-DOPA treatment. Interactions between drug and transition, $F(1,17) = 4.09$, $p = 0.06$, or drug and reward (which would indicate a drug-induced change in model-free control), $F(1,17) = 1.10$, $p = 0.31$, were not significant.

Figure 2B shows the difference in stay probability between drug states corrected for a main effect of drug. Note that dopamine treatment particularly affected choices after unrewarded trials and a post hoc contrast; testing for a differential drug effect after unrewarded compared to rewarded trials confirmed this was significant, $F(1,17) = 12.68$, $p = 0.002$. Figures 2C–2F illustrate how a number of hypothesized effects of L-DOPA might manifest itself in a stay-switch analysis (see Figure S1 available online for a validation of these hypotheses using computational modeling). Qualitatively, the data in Figure 2B resemble a shift toward model-based control, most notable after unrewarded trials. In contrast, our results do not resemble any of the putative model hypotheses that invoke modulation of a model-free system.

Given the broad effects of drug in this analysis, we next employed computational modeling to provide an in-depth understanding of this pharmacological effect. The value of using such an approach is that a stay-switch analysis only considers variables on trial $n - 1$, while a computational model encompasses an integration over a longer reward history and attributes any behavioral change to a specific computational process.

Model comparisons (Table S2) between a fully parameterized hybrid model (Daw et al., 2011; Gläscher et al., 2010) and various reduced nested versions favored a model with the parameters learning rate α, softmax temperature β, perseverance π, and relative degree of model-based/model-free control ω as best fit. We then fitted parameters individually for each

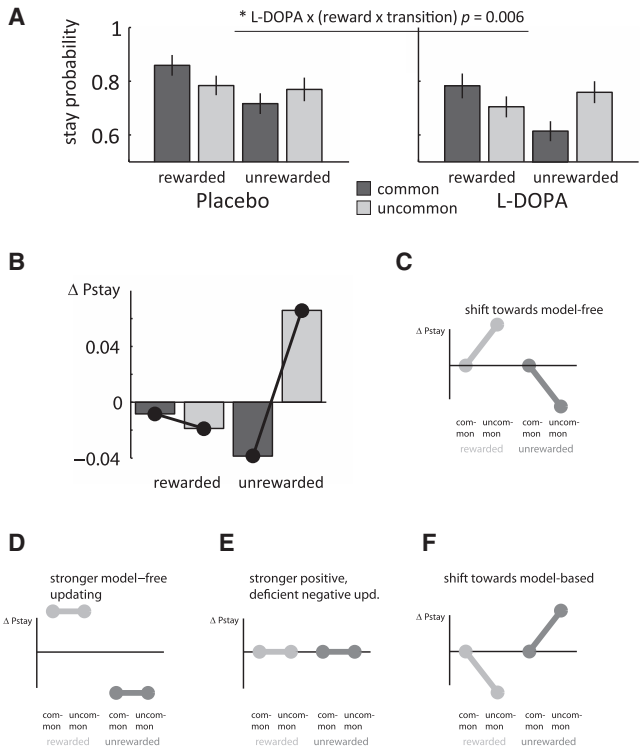

Figure 2. Results Stay-Switch Analysis

(A) Subjects' task behavior showed characteristics of both model-free and model-based influences, demonstrating that subjects combined both strategies in the task. The reward × transition interaction (a measure of the extent to which subjects consider the task structure) was significantly larger in L-DOPA compared to placebo, indicating stronger model-based behavior. Error bars represent SEM. (B) Difference in stay probability between L-DOPA and placebo condition, corrected for the main effect of drug. The observed interaction indicates a shift toward model-based choice (see F) but shows no resemblance to any of the three effects implicating the model-free system (see C–E). (C–F) Illustration of expected differences in stay probability for hypothetical drug effects. See Figure S1 and Table S1 for validation of these hypotheses. (C) Trials after uncommon transitions (second and fourth bar) are discriminatory between model-free and model-based choice, whereas both models make equal predictions for trials after common transitions (cf. Figure 1C). A shift toward model-free control would be indicated by an increased propensity to stay with the chosen pattern after uncommon rewarded trials and an increase in switching after uncommon unrewarded trials. (D) Stronger or faster model-free learning would increase the reward-dependent effect and be expressed as a general increase to stay after rewarded trials and general decrease to stay after unrewarded trials. (E) A selective enhancement of positive updating paired with impairment in negative updating might not change mean-corrected stay probabilities. This is because enhanced positive updating leads to a stronger propensity to stay after rewarded trials, while impaired updating of unrewarded trials decreases the propensity to switch after such trials. (F) Opposite to (C), a shift toward model-based control is expressed by enhanced sensitivity to the task structure.

subject and drug state after applying logistic or exponential transformations to bounded model parameters (α, β, π, ω) to gain Gaussian distributed fitted parameter values (a, b, p, w), permitting the use of parametric tests for differences between sessions. All reported p values are from two-tailed paired t tests.

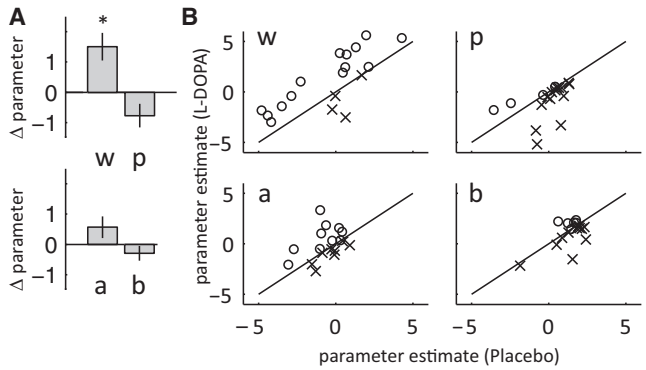

Figure 3. Results Computational Learning Model

(A) Group-averaged parameter differences in the computational learning model between L-DOPA and placebo. Parameter w represents a measure of model-based over model-free control. Error bars represent SEM. (B) Single-subject data for parameter values in (A). Each data point represents the parameter value of a single subject. Subjects above the diagonal (circles) had higher parameter values in the L-DOPA compared to placebo condition, while subjects below (crosses) had smaller parameter values. The relative degree of model-based control was higher in the L-DOPA condition in 14 out of 18 subjects. See Figure S1 and Table S1 for validation of the winning model and Table S2 for model comparison details.

In line with the stay-switch results, we found a significant increase in the model-based weighting parameter w, p = 0.005, (positive in 14 out of 18 subjects) and a trend-level decrease in the perseverance parameter p, p = 0.06, under L-DOPA compared to placebo. Learning rate a, p = 0.45, and softmax temperature b, p = 0.34, did not differ between drug states (Figure 3). We note that, overall, fitted parameter values were in a similar range as those in Daw et al. (2011) (Table 1). As model-based choice is superior to model-free choice in this task, we found a significant positive correlation between subjects' relative degree of model-based control (w) and total earnings, r = 0.4, p = 0.01 (Figure S2). There was no evidence for differences in drowsiness or general alertness (Bond et al., 1974) between sessions (paired t tests over each score; smallest p > 0.1) or in average response times between drug states (first stage $RT_{L\text{-}DOPA}$ = 593 ms, $RT_{Placebo}$ = 586 ms; paired t test, p = 0.70).

Note that in the preceding analysis we employed the same computational models as the authors in the original study utilizing this task (Daw et al., 2011). We also constructed additional computational models to further explore the observed shift in control and to examine whether dopamine asserts its effect predominantly on the model-free or model-based system. Some studies have suggested that dopamine levels might have differential effects on positive and negative updating (Frank et al., 2004; Pessiglione et al., 2006). We therefore tested a model with separate learning rates for positive (a+) and negative (a−) updating. The learning rates were not significantly different between L-DOPA and placebo (paired t test: a+, p = 0.52 and a−, p = 0.43). The use of the same values at the second stage for both model-free and model-based systems ignores evidence that model-based and model-free learning use different neural structures (Balleine and O'Doherty, 2010;

Table 1. Best-Fitting Parameter Estimates, Shown Separately for Both Drug Conditions as Median and Quartiles across Subjects

	α	β	π	ω
Placebo				
25th percentile	0.25	3.4	0.63	0.07
Median	0.45	5.4	1.37	0.58
75th percentile	0.57	7.3	2.20	0.79
L-DOPA				
25th percentile	0.25	1.8	0.28	0.14
Median	0.37	4.7	0.70	0.78
75th percentile	0.59	7.7	1.50	0.95

Wunderlich et al., 2012) and, as such, might learn the second-stage values separately. To test this, we implemented a model containing separate representations of second-stage values and learning rates for the model-based and model-free system. The model-based learning rate was higher than the model-free learning rate ($p = 0.001$). However, concurring with the results from our original computational implementation, there was no change in either learning rate with drug condition (α model-free $p = 0.33$, model-based $p = 0.76$). An alternative computational implementation of model-free RL, the actor-critic model, learns values and action policies separately (Sutton and Barto, 1998). To test whether L-DOPA might alter updating of action policies rather than impacting on value updating, we implemented a hybrid model in which the original model-free component was replaced with an actor-critic component. In line with the absence of a significant difference in the parameters of the original model-free implementation, this analysis did not show any significant difference between drug states in either the learning parameter α ($p = 0.17$) for state value or η for policy updating ($p = 0.51$).

Finally, we tested for order effects by repeating the analyses with session instead of drug as factor. There were no significant differences in either stay-switch behavior (repeated-measures ANOVA; main effect of session $F(1,17) < 1$; session × reward, $F(1,17) < 1$; session × (reward × transition), $F(1,17) = 1.37$, $p = 0.26$) or parameter fits in the computational analysis with session as a grouping factor (two-tailed paired t tests; *a*: $p = 0.15$; *b*: $p = 0.31$; *p*: $p = 0.97$; *w*: $p = 0.37$). Thus, our results provide compelling evidence for an increase in the relative degree of model-based behavioral control under conditions of elevated dopamine.

DISCUSSION

It is widely believed that both model-free and model-based mechanisms contribute to human choice behavior. In this study, we investigated a modulatory role of dopamine in the arbitration between these two systems and provide evidence that L-DOPA increases the relative degree of model-based over model-free behavioral control.

The use of systemic L-DOPA combined with a purely behavioral approach precludes strong conclusions about the precise physiological underpinnings of the observed shift to model-based control. Nevertheless, we provide a number of possible explanations for how this effect might be mediated in the brain that could guide further studies. First, increased dopamine levels may improve performance of component processes of a model-based system. Dopamine has previously been associated with an enhancement of cognitive functions such as reasoning, rule learning, set shifting, planning, and working memory (Clatworthy et al., 2009; Cools and D'Esposito, 2011; Cools et al., 2002; Lewis et al., 2005; Mehta et al., 2005), and these processes are most likely coopted during model-based decisions. Previous theoretical considerations link a system's performance to its relative impact on behavioral control, such that the degree of model-based versus model-free control depends directly on the relative certainties of both systems (Daw et al., 2005). Increased processing capacity might enhance certainty in the model-based system and would thus predict the observed shift in behavioral control that we detail here.

Second, a more conventional account is that increased dopamine exerts its effect through an impact on a model-free system. According to this view, excessive dopamine disrupts model-free reinforcement learning, which is then compensated for by increased model-based control. Specifically, elevated tonic dopamine levels may reduce the effectiveness of negative prediction errors (Frank et al., 2004; Voon et al., 2010). However, this explanation fails to account for the results presented here. First, a disruption of negative prediction errors under L-DOPA would change stay probabilities independent of transition type (Figure 2E), which is incompatible with the drug × reward × transition interaction observed here (Figure 2B). Second, any such model-free impairment would have impacted learning of second-stage values (which in this task are assumed to be learnt via prediction errors irrespective of the control on the first stage; Daw et al., 2011) and manifested in noisier choices or altered learning rates. We did not observe such an effect on the softmax temperature *b* or learning rate *a*. This effect was still absent when we fit alternative models employing separate learning rates and temperatures for the first and second stage or separate learning rates for positive and negative updating. Together, this argues against the idea that L-DOPA in our study enhanced the relative degree of model-based behavior through a disruption of the model-free system.

Finally, dopamine could facilitate switching from one type of control to the other akin to the way it decreases behavioral persistence (Cools et al., 2003). It is known that over the course of instrumental learning, the habitual system assumes control from the goal-directed system (Adams, 1982; Yin et al., 2004), but the goal-directed system can quickly regain control in unforeseen situations (Isoda and Hikosaka, 2011; Norman and Shallice, 1986). This could explain why we observe a stronger switch to model-based behavior after unrewarded trials: the lack of rewarding feedback may prompt the need to reevaluate available options and invest more energy to prevent another nonrewarding event by switching to model-based control. Note that it is possible and indeed likely that a facilitation of control switching under L-DOPA works in concert with an enhancement of the model-based system itself.

The predominant view in computational and systems neuroscience holds that phasic dopamine underlies model-free behavior by encoding reward prediction errors. On the other

hand, animal and cognitive approaches emphasize a role for dopamine in model-based behavior such as planning and reasoning (Berridge, 2007; Clatworthy et al., 2009; Cools and D'Esposito, 2011; Robbins and Everitt, 2007). Contrasting with interest in the model-free and model-based system separately is the lack of data on the arbitration between these two behavioral controllers. Our experiment fills this gap by pitting model-free and model-based control against each other in the same task and in so doing provides strong evidence for an involvement of dopamine in the arbitration between model-free and model-based control over behavior.

Our findings advocate an effect of L-DOPA on the arbitration between model-based and model-free control, without a modulation of the model-free system itself. Note that the majority of studies reporting enhanced or impaired learning under dopaminergic drugs used either Parkinson's disease (PD) patients (Frank et al., 2004; Voon et al., 2010) or involved agents that primarily act at D2 receptors (Cools, 2006; Frank and O'Reilly, 2006). In contrast with these studies, we did not find evidence for any modulation by L-DOPA of model-free learning rates or indeed evidence of impaired model-free choices. These deviations might partly be explained by PD patients' more severely reduced dopamine availability off their dopamine replacement therapy (in contrast to our placebo condition) and the much higher doses of medication involved in PD treatment. Consistent with this explanation is that the effect of L-DOPA on instrumental learning in healthy volunteers was found to be significant only when compared to an inhibition of the dopamine system (via haloperidol) but not when compared to placebo (Pessiglione et al., 2006).

Our task does not allow us to dissociate between learning and performance effects. Previous work has suggested interactions between model-based and model-free systems during learning. In this framework, reward prediction errors that are in line with model-based predictions are enhanced, while reward prediction errors that are in opposition with model-based predictions are attenuated (confirmation bias) (Biele et al., 2011; Doll et al., 2009, 2011). In support of this, neuroimaging findings based on the present task showed evidence that ventral striatal BOLD at the time of feedback, typically associated with prediction error signals, contained a model-based component (Daw et al., 2011). This raises the possibility that model-free and model-based systems are not segregated systems whose influence is weighted at the time of choice. Instead, choices could also be made by a model-free system in which learning is modulated by transition probabilities. In this study, we cannot unambiguously differentiate between these accounts and further fine-grained investigations, in part motivated by the present data, are required to understand this complex issue.

Dopamine itself is a precursor to norepinephrine and epinephrine, potentially contributing to the observed effects. However, L-DOPA administration causes a linear increase in dopamine levels in the brain without affecting norepinephrine levels (Everett and Borcherding, 1970). Another possibility would be that L-DOPA exerts effects through interactions with the serotonin system. Such an interaction, between dopamine and serotonin, is known to play a role in a range of higher-level cognitive functions (Boureau and Dayan, 2011).

By implicating dopamine in behavioral control, we open the door to further experiments aimed at elucidating the precise neural mechanisms underlying the arbitration between both controllers. While theoretical considerations afford a number of ways for how this arbitration might be implemented in the brain (Daw, 2011; Keramati et al., 2011), our results provide empirical evidence that dopamine influences the relative degree between model-free and model-based control.

EXPERIMENTAL PROCEDURES

Subjects

Eighteen healthy males (mean age: 23.3 [SD: 3.4]) participated in two separate sessions. Data from two additional subjects were not included in the analysis as those subjects misunderstood instructions and performed at chance level. The UCL Ethics committee approved the study and subjects gave written informed consent before both sessions.

Dopamine Drug Manipulation

Subjects were tested in a double-blind, fully counterbalanced, repeated-measures setting on L-DOPA (150 mg $_{L}$-3,4-dihydroxyphenylalanine / 37.5 mg benserazide; Madopar, Roche) and on placebo (500 mg calcium carbonate; Calcit, Procter and Gamble) dispersed in orange squash. The task was administered 55.0 (SD: 4.7) min after drug administration. Sessions one and two were approximately 1 week apart (at least 4, but no more than 14 days), with both sessions at the same time of day. All subjects except one participated in the morning to minimize time-of-day effects. We assessed drug effects on self-reported mental state using a computerized visual analog scale immediately before starting the task (Bond et al., 1974).

Task

We drew on Daw et al. (2011)'s two-step choice task to assess the relative degree of model-based versus model-free decision making. Our version of the task was identical to Daw et al.'s except for different stimulus images (semantically irrelevant fractals), a slightly larger dynamic range of reward probabilities, and more rapid trial timings. Subjects completed 201 trials and were given a break after trials 67 and 134. Please see Supplemental Experimental Procedures for full task description.

Stay-Switch Behavior

Stay probabilities at the first stage (the probability to choose the same stimulus as in the preceding trial), conditional on transition type of the previous trial (common or uncommon), reward on the previous trial (reward or no reward), and drug state (L-DOPA or placebo) were entered into a three-way repeated-measures ANOVA.

Computational Modeling

We fit a previously described hybrid model (Gläscher et al., 2010; Daw et al., 2011) to choice behavior. This model contains separate terms for model-free and model-based stimulus values at the first stage. These values are weighted by a parameter w to compute an overall value for each stimulus. The first-stage choice is then made using a softmax function dependent on relative stimulus values and the subject's choice on the previous trial. For a full description of the model, see Supplemental Experimental Procedures.

Model Fitting

We used a hierarchical model-fitting strategy, which takes into account the likelihood of individual subject choices c_i given the individual subject parameters a_i, b_i, p_i, w_i and also the likelihood of the individual subject parameters given the parameter distribution in the overall population across conditions. This regularizes individual subjects' parameter fits, rendering them more robust toward overfitting.

This two-stage hierarchical procedure is a simplified estimation strategy of the iterative expectation – maximization (EM) algorithm (see Supplemental

Experimental Procedures for details, and for an in-depth discussion see also Daw, 2011).

Importantly, our main results are independent of the parameter regularization: the weighting parameter w was significantly ($p = 0.02$) higher in the L-DOPA condition compared to placebo, even when testing individual subject parameters from the maximum likelihood fit during the first step.

Parameter Covariance

Covariance between parameters would indicate that two parameters might be redundant, potentially rendering parameter values more difficult to interpret. There were no significant pairwise correlations between any of our parameters across subjects (paired t tests: all individual $p > 0.05$).

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, two tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2012.03.042.

ACKNOWLEDGMENTS

We thank Tamara Shiner for help with drug administration. We are also grateful to Peter Dayan, Roshan Cools, Marc Guitart-Masip, and Quentin Huys for helpful comments on the manuscript. This study was supported by the Wellcome Trust (Ray Dolan Programme Grant number 078865/Z/05/Z; Peter Smittenaar 4 year PhD studentship; The Wellcome Trust Centre for Neuroimaging is supported by core funding 091593/Z/10/Z) and Max Planck Society.

Accepted: March 22, 2012
Published: August 8, 2012

REFERENCES

Adams, C.M. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q. J. Exp. Psychol. *34B*, 77–98.

Balleine, B.W., and O'Doherty, J.P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology *35*, 48–69.

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

Biele, G., Rieskamp, J., Krugel, L.K., and Heekeren, H.R. (2011). The neural basis of following advice. PLoS Biol. *9*, e1001089.

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C.E., Daw, N., Dibó, G., Takáts, A., Bereczki, D., and Gluck, M.A. (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.

Bond, A.J., James, D.C., and Lader, M.H. (1974). Physiological and psychological measures in anxious patients. Psychol. Med. *4*, 364–373.

Boureau, Y.L., and Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. Neuropsychopharmacology *36*, 74–97.

Clatworthy, P.L., Lewis, S.J., Brichard, L., Hong, Y.T., Izquierdo, D., Clark, L., Cools, R., Aigbirhio, F.I., Baron, J.C., Fryer, T.D., and Robbins, T.W. (2009). Dopamine release in dissociable striatal subregions predicts the different effects of oral methylphenidate on reversal learning and spatial working memory. J. Neurosci. *29*, 4690–4696.

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., and D'Esposito, M. (2011). Inverted-U-shaped dopamine actions on human working memory and cognitive control. Biol. Psychiatry *69*, e113–e125.

Cools, R., Stefanova, E., Barker, R.A., Robbins, T.W., and Owen, A.M. (2002). Dopaminergic modulation of high-level cognition in Parkinson's disease: the role of the prefrontal cortex revealed by PET. Brain *125*, 584–594.

Cools, R., Barker, R.A., Sahakian, B.J., and Robbins, T.W. (2003). L-Dopa medication remediates cognitive inflexibility, but increases impulsivity in patients with Parkinson's disease. Neuropsychologia *41*, 1431–1441.

Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Affect, Learning and Decision Making. Attention and Performance, E. Phelps, T. Robbins, and M. Delgado, eds. (Oxford: Oxford University Press).

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

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215.

Dickinson, A., and Balleine, B.W. (2002). The role of learning in the operation of motivational systems. In Stevens' Handbook of Experimental Psychology, H. Pashler and R. Gallistel, eds. (New York: John Wiley & Sons), pp. 497–533.

Doll, B.B., Jacobs, W.J., Sanfey, A.G., and Frank, M.J. (2009). Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res. *1299*, 74–94.

Doll, B.B., Hutchison, K.E., and Frank, M.J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J. Neurosci. *31*, 6188–6198.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. *12*, 961–974.

Everett, G.M., and Borcherding, J.W. (1970). L-DOPA: effect on concentrations of dopamine, norepinephrine, and serotonin in brains of mice. Science *168*, 847–850.

Frank, M.J., and O'Reilly, R.C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav. Neurosci. *120*, 497–517.

Frank, M.J., Seeberger, L.C., and O'Reilly, R.C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science *306*, 1940–1943.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595.

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

Isoda, M., and Hikosaka, O. (2011). Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behaviors. Eur. J. Neurosci. *33*, 2058–2069.

Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput. Biol. *7*, e1002055.

Lewis, S.J., Slabosz, A., Robbins, T.W., Barker, R.A., and Owen, A.M. (2005). Dopaminergic basis for deficits in working memory but not attentional set-shifting in Parkinson's disease. Neuropsychologia *43*, 823–832.

Mehta, M.A., Gumaste, D., Montgomery, A.J., McTavish, S.F., and Grasby, P.M. (2005). The effects of acute tyrosine and phenylalanine depletion on spatial working memory and planning in healthy volunteers are predicted by changes in striatal dopamine levels. Psychopharmacology (Berl.) *180*, 654–663.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

Norman, D., and Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In Consciousness and Self-Regulation. Advances in Research and Theory, R. Davidson, G. Schwartz, and D. Shapiro, eds. (New York, London: Plenum Press), pp. 1–18.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M.C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M.R., and Obeso, J.A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. Nat. Rev. Neurosci. *11*, 760–772.

Robbins, T.W., and Everitt, B.J. (2007). A role for mesencephalic dopamine in activation: commentary on Berridge (2006). Psychopharmacology (Berl.) *191*, 433–437.

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

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (Cambridge, MA: MIT Press).

Voon, V., Pessiglione, M., Brezing, C., Gallea, C., Fernandez, H.H., Dolan, R.J., and Hallett, M. (2010). Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. Neuron *65*, 135–142.

Wunderlich, K., Dayan, P., and Dolan, R.J. (2012). Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci. *15*, 786–791.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. *19*, 181–189.

Neuron, Volume *75*

Supplemental Information

Dopamine Enhances Model-Based over Model-Free Choice Behavior

Klaus Wunderlich, Peter Smittenaar, and Raymond J. Dolan

Figure S1

To confirm that our winning model can capture our key behavioral findings (i.e. the drug*reward*transition interaction on stay-switch behavior) we generated data for 500 virtual subjects on this task using the best-fitting parameters. These data were then subjected to a stay-switch analysis. We found an identical pattern of effects in these generated data as observed empirically in our participants (Figure S1A). Most importantly, the data generated by the model showed a significant 3-way interaction, indicating that the model indeed captures key components of the data (see Table S1). Note that, as expected, the model did not replicate the asymmetry in rewarded versus unrewarded trials shown in Figure 2B.

The idealized hypotheses put forward for the stay-switch analysis in Figures 2C-2F were based on ideas derived from previous literature. To validate these hypotheses we generated choices for virtual subjects, but now with adjustments to parameters based on specific hypotheses (Figures S1B-S1D). Our key hypotheses are fully supported by these simulations, showing that the computational models capture the key behavioral signatures of model-free and model-based behavior.

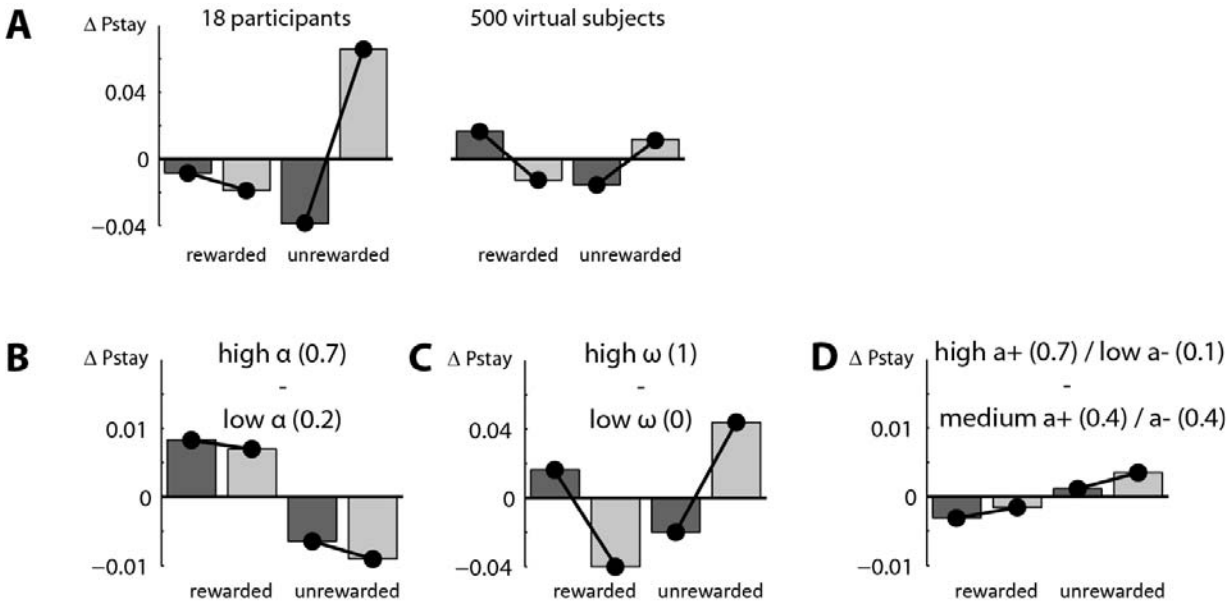

Figure S1. Validation of model and hypotheses (related to Figure 2). (**A**) Mean-corrected $P(stay)_{ON}$ - $P(stay)_{OFF}$ for 18 participants reported in the study (left) and 500 virtual subjects using best-fitting parameters in the winning model (right). (**B**) Modulation of the model-free learning rate α. A change in learning rate alters stay probability after rewarded versus unrewarded trials, but does not interact with transition. This is equivalent to Figure 2D in main text. (**C**) Model-based (ω = 1) versus model-free agent (ω = 0) shows a stronger reward*transition interaction. This is equivalent to Figure 2F in the main text. (**D**) Increase in positive learning rate and decrease in negative learning rate does not change relative stay probabilities, similar to our prediction in Figure 2E.

Figure S2

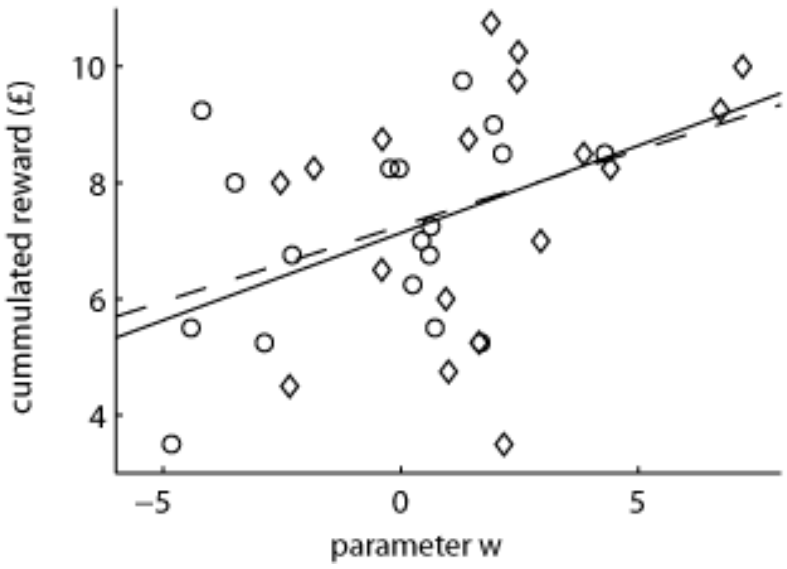

Figure S2. Task performance increases with degree of model-based control (related to Figure 3). Performance-based reward per session (£) correlated with degree of model-based control as indicated by the parameter fit *w* ($r = 0.40$, $p = 0.01$). This relation was still significant even when we control for the 3 other parameter values using partial correlations ($r_{wr} = 0.34$, $p = 0.04$). When we test for the correlation within each session, we find very similar regression coefficients for L-DOPA and placebo albeit each individual test is not significant due to the reduced statistical power (for placebo: $r=0.4$, $p=0.10$; for L-DOPA: $r=0.39$, $p=0.12$). Data points represent individual subjects and session (diamond: L-DOPA, circle: placebo). Regression lines plotted separately for L-DOPA (solid) and placebo (dashed).

Table S1. Statistical comparison of model-generated versus participant data (related to Figure 2)

Effect	18 participants		500 virtual subjects	
	F(1,17)	p	*F*(1,499)	p
drug	7.04	= .02	83.00	< .001
reward	23.30	< .001	6.01	= .02
transition	< 1	~	< 1	~
drug x reward	1.10	= .31	< 1	~
drug x transition	4.09	= .06	< 1	~
reward x transition	9.75	= .006	561.79	< .001
drug x reward x transition	9.86	= .006	16.62	< .001

The data generated by the model in Figure S1A was subjected to the same ANOVA as the participant data. The stay-switch data generated by the model showed the same effects as found in participants, most notably the three-way interaction that supports our claim that L-DOPA enhances model-based behavior. The model thus provides a reasonable account of the data. Identical patterns exist between the two datasets, given the statistical model used. Highlighting indicates significant effects.

Table S2. Model comparison (related to Table 1)

A. BIC scores

Model parameters	BIC	# parameters
$\alpha1, \alpha2, \beta1, \beta2, \lambda, \pi, \omega$	7286	7
$\alpha1, \alpha2, \beta1, \beta2, \pi, \omega$	7251	6
$\alpha, \beta1, \beta2, \pi, \omega$	7109	5
$\alpha1, \alpha2, \beta, \pi, \omega$	7160	5
$\alpha, \beta, \lambda, \pi, \omega$	7164	5
$\alpha+, \alpha-, \beta, \pi, \omega$	7192	**5**
$\alpha, \beta, \pi, \omega$	**7097**	**4**
α, β, ω	7846	3
α, β	8221	2
MF/MB learning rates	7018	5
Actor/critic learning	7308	5

B. Bayesian Model Comparison

	Placebo		L-DOPA	
Alternative model to $\alpha, \beta, \pi, \omega$	Better in #subjects	Exceedance probability	Better in #subjects	Exceedance probability
$\alpha1, \alpha2, \beta1, \beta2, \lambda, \pi, \omega$	17	>0.999	15	0.999
$\alpha1, \alpha2, \beta1, \beta2, \pi, \omega$	14	0.997	15	0.999
$\alpha, \beta1, \beta2, \pi, \omega$	13	0.970	14	0.998
$\alpha1, \alpha2, \beta, \pi, \omega$	15	>0.999	15	>0.999
$\alpha+, \alpha-, \beta, \pi, \omega$	12	>.831	15	>.996
$\alpha, \beta, \lambda, \pi, \omega$	16	>0.999	17	>0.999
α, β, ω	16	>0.999	12	0.944
α, β	16	>0.999	12	0.911
MF/MB learning rates	16	0.999	14	0.999
Actor/critic learning	18	>0.999	17	>0.999

A model with the parameters learning rate, softmax temperature, perseverance, and model-based weight fitted subjects' choices best in a model comparison that considers differences in model complexity. More complex model variants included separate model-free RL parameters for the first and second stage, and eligibility traces. A: the BIC scores are the Bayesian equivalent to a fixed effects analysis. B: the exceedance probability corresponds to a mixed effects analysis (Stephan et al., 2009).

We calculated posterior model probabilities for each subject and the group of subjects. In brief, the procedure by Stephan et al. rests on treating the model as a random variable and estimating the parameters of a Dirichlet distribution, which describes the probabilities for all models considered. These probabilities then define a multinomial distribution over model space, allowing one to compute how likely it is that a model generated the subjects' data. To decide which model is more likely, we use the conditional model probabilities to quantify an exceedance probability, i.e. a belief that a particular model is more likely than the other model, given the group data. Note that even though the average BIC of the model with separate representations of model-free/model-based (MF/MB) second stage values is slightly lower than the α, β, π, ω model, random effects analysis shows that behavior in the majority of our subjects and data at the population level can be much better ($p>0.999$) explained by the α, β, π, ω model. We therefore used this model for all analyses displayed in the figures.

Supplemental Experimental Procedures

Detailed task description

Each trial consisted of two stages, both requiring a choice between two stimuli. Each choice option was represented by a fractal in a colored box on a black background (Figure 1). At every choice, subjects had to respond within two seconds using the left/right cursor keys or the trial was aborted. Subjects rarely missed a trial [mean: 0.4%, SD: 1.5%], and those missed trials were omitted from analysis. Choice at the first stage always involved the same two stimuli left/right randomized. After subjects made their response the rejected stimulus disappeared from the screen and the chosen stimulus moved to the top of the screen. After 1.5 seconds one of two second stage stimulus pairs appeared, with the transition from first to second stage following fixed transition probabilities. Each first stage option was more strongly (with a 70% transition probability) associated with one of the two second stage pairs, a crucial factor in allowing us to distinguish model-free from model-based behavior (see below). After the second choice, the chosen option remained on the screen, together with a reward symbol (a pound coin) or a 'no reward' symbol (a red cross). Each of the 4 stimuli in stage 2 had a reward probability between 0.2 and 0.8. Those reward probabilities drifted slowly and independently for each of the four second stage options in every trial through a diffusion process with Gaussian noise (mean 0, SD 0.025).

Prior to the experiment, subjects were given explicit information about the task structure; namely that for each stimulus on the first stage one of the two transition probabilities was higher than the other, and that these transition probabilities remained constant throughout the experiment. Subjects were also told that reward probabilities on the second stage were independent of each other and would change slowly over time. To minimize the variance resulting from different outcome sequences we used the same two templates for outcome probabilities on all subjects. The assignment of templates to session and drug state was fully counterbalanced across subjects. On both days, subjects practiced 50 trials with different stimuli and outcome probabilities before starting the task. The main task consisted of 201 trials with short breaks after trial 67 and 134. Subjects' payout was related to a flat amount plus their overall accumulated rewards from both sessions (total range 16-30.40 in £s).

Computational modeling

In the following we denote the model-free value V_{s1}^{MF} and the model-based value V_{s1}^{MB} for first stage stimuli $s1 \in [1,2]$. The hybrid model computes the actual value that is used in determining choice as weighted linear combination

$$V_{s1}^{Hybrid} = \omega * V_{s1}^{MB} + (1-\omega) * V_{s1}^{MF}. \quad (1)$$

Values for the four stimuli at the second stage (stimuli $s2 \in [3..6]$) are updated identically for both models according to reward prediction errors (Rummery and Niranjan, 1994):

$$V_{s2}(t+1) = V_{s2}(t) + \alpha_2 (r - V_{s2}(t)). \quad (2)$$

At the first stage, model-free 'cached' values are updated according to temporal difference learning with reward prediction errors and eligibility traces:

$$V_{s1}^{MF}(t+1) = V_{s1}^{MF}(t) + \alpha_1(V_{s2\_chosen}(t) - V_{s1}^{MF}(t)) + \lambda\alpha_1(r - V_{s1}^{MF}(t)), \quad (3)$$

where α_1 / α_2 are learning rates at the first and second stage, and λ is a gain parameter for the eligibility traces.

Model-based values are calculated anew for each and every trial in a forward looking manner by multiplying the state values of the better option at the second stage with the state transition probabilities:

$$V_1^{MB} = 0.7*\max(V_3, V_4) + 0.3*\max(V_5, V_6) \text{ and } V_2^{MB} = 0.3*\max(V_3, V_4) + 0.7*\max(V_5, V_6). \quad (4)$$

Based on simulations by the authors of the original task we similarly simplified model-based learning by the premise that learning of state transitions quickly converges to stable values and hence we did not update transition probabilities by explicitly modeling state prediction errors (see supplementary materials in Daw et al. (2011) for a comprehensive discussion of this matter).

The probability *P* of choosing stimulus 1 (in a choice between stimulus 1 with value *V1* and stimulus 2 with value *V2*) was computed in stage 1 according to a softmax choice function dependent on the relative stimulus values and choice *C* in the previous trial.

$$P(1) = 1 / (1 + \exp(-\beta_1(V1 - V2) - \pi(C1 - C2))) \quad (5)$$

and similarly in stage 2

$$P(1) = 1 / (1 + \exp(-\beta_2(V1 - V2))) \quad (6)$$

For additional in-depth information on task and computational model see also Daw et al. (2011).

A model comparison using BIC scores (see below) of the full model with various reduced versions indicated that the best fitting model in the present experiment was a reduced model with single learning and softmax parameters for both stages, without the eligibility term (Table S2). This model includes variables for learning rates $\alpha_{1/2}$, inverse softmax temperatures $\beta_{1/2}$, perseverance π, and a parameter ω for the relative degree of model-based versus model-free control. Each of these parameters represents different aspects of choice behavior. The learning rate α captures the extent to which new information at outcome is used for learning, i.e. the learning speed; β measures the discriminability between two options, with a larger value pertaining to more precise choices when the values of alternative options are relatively close together; the persistency π is an index of the tendency to choose the same option as in the previous trial regardless of value (Kable and Glimcher, 2007), and parameter ω represents the extent to which one or other system drives a participant's behavior. By comparing ω across drug states we were able to examine how L-DOPA changes the relative importance of the model-free and model-based system in driving behavior in this task.

We applied logistic/exponential transformations before fitting parameters to transform bounded parameters into Gaussian distributed parameter values $x_i \sim N(\mu_x, \sigma_x)$, with population mean μ_x and

standard deviation σ_x. This transformation is justified by the premise that each individual subject (with parameter value x_i) is randomly drawn from a population of subjects with normally distributed parameters (with population mean μ_x and standard deviation σ_x) (Daw, 2011). It is important for our analysis because normally distributed parameter values permit the use of parametric tests and random effects statistics.

We transformed [0,1]-bounded α and ω into a Gaussian scale using the logistic function

$$\alpha = 1/(1+\exp(-a)), \text{ and } \omega = 1/(1+\exp(-w)) \quad (7)$$

and the logarithmically scaled β and π using the exponential function

$$\beta = \exp(b), \text{ and } \pi = \exp(p). \quad (8)$$

We denote model parameters by Greek letters and the Gaussian transformation by their respective Latin letters.

Hierarchical model fitting

A subject *i* drawn from the population has a set of parameters (i.e. a_i, b_i, p_i, w_i) according to a statistical distribution that characterizes the distribution of parameters in the population (with means μ_a, μ_b, μ_p, μ_w and standard deviations σ_a, σ_b, σ_p, σ_w). Adopting a model of the parameters in the population gives us a two-level hierarchical model of how a full dataset is produced. Each subject's parameters are drawn from population distributions, then the choice values and the observable choice data are generated according to the RL model with those parameters.

In a first pass we fitted parameters to every individual subject by maximizing the likelihood of subjects' choices given the parameterized model:

$$\begin{aligned} &L = P\left(c_i \mid \mu_a,\mu_b,\mu_p,\mu_w,\sigma_a,\sigma_b,\sigma_p,\sigma_w\right) = \\ &\int da_i db_i dp_i dw_i P(c_i \mid a_i,b_i,p_i,w_i) P(a_i \mid \mu_a,\sigma_a) P(b_i \mid \mu_b,\sigma_b) P(p_i \mid \mu_p,\sigma_p) P(w_i \mid \mu_w,\sigma_w) \end{aligned} \quad (9)$$

We next estimated mean and variance of the parameter distribution in the population based on our subject sample (e.g. $\mu_a = 1/N \sum_i^N a_i$ and $\sigma_a = \sqrt{\frac{1}{N}\sum_i^N \left(a_i - \mu_a\right)^2}$).

In a third step we refitted single subject parameter values by maximizing over both the likelihood of subjects' choices given the parameters and the likelihood for individual subject parameter values given the distribution of parameters in the population:

$$\begin{aligned} &P\left(a_i,b_i,p_i,w_i \mid c_i,\mu_a,\mu_b,\mu_p,\mu_w,\sigma_a,\sigma_b,\sigma_p,\sigma_w\right) \\ &\propto P\left(c_i \mid a_i b_i p_i w_i\right) \times P\left(a_i b_i p_i w_i \mid \mu_a,\mu_b,\mu_p,\mu_w,\sigma_a,\sigma_b,\sigma_p,\sigma_w\right) \end{aligned} \quad (10)$$

Model comparison

We performed model-selection between a fully parameterized hybrid model (accounting for learning rates and softmax choice temperatures separately at the first and second stage, and allowing for eligibility trace updating, perseverance, and a model-free versus model-based weighting parameter) and various reduced versions of this model by fitting the free parameters of each model across both sessions. Comparing Bayesian Information Criterion (BIC) (Schwarz, 1978), which considers differences in model complexity, we found a best fit for the model with a common learning rate and temperature for both stages, and a perseverance and weight parameter. We calculated BIC as

$$\mathrm{BIC} = 2L + k\ln(n) \qquad (11)$$

where L is the negative log likelihood function, n the number of choices and k the number of free model parameters.

Supplemental References

Balleine, B.W., and O'Doherty, J.P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology *35*, 48-69.

Biele, G., Rieskamp, J., Krugel, L.K., and Heekeren, H.R. (2011). The neural basis of following advice. PLoS Biol *9*, e1001089.

Daw, N.D. (2011). Trial-by-trial data analysis using computational models. In Affect, Learning and Decision Making. Attention and Performance., E. Phelps, T. Robbins, and M. Delgado, eds. (Oxford: Oxford University Press).

Daw, N.D., Gershman, S.J., Dayan, P., Seymour, B., and Dolan, R.J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron *69*, 1204-1215.

Doll, B.B., Hutchison, K.E., and Frank, M.J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. The Journal of neuroscience : the official journal of the Society for Neuroscience *31*, 6188-6198.

Frank, M.J., Seeberger, L.C., and O'Reilly R, C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science *306*, 1940-1943.

Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Nat Neurosci *10*, 1625-1633.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042-1045.

Rummery, G.A., and Niranjan, M. (1994). On-Line Q-Learning Using Connectionist Systems. In CUEF/F-INFENG/TR.

Schwarz, G.E. (1978). Estimating the dimension of a model. Annals of Statistics *6*, 461-464.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. Neuroimage *46*, 1004-1017.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (MIT Press, Cambridge, MA).

Wunderlich, K., Dayan, P., and Dolan, R.J. (2012). Mapping value based planning and extensively trained choice in the human brain. Nature neuroscience *15*, 786–791.