Neuron, Volume *83*

Supplemental Information

A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice

Jeffrey Cockburn, Anne G.E. Collins, and Michael J. Frank

Supplemental Data

Justification of gene selection

The DARPP-32 gene is associated with an intracellular protein that is highly concentrated in striatum. When phosphorylated by D_1 receptor stimulation, protein photophase-1 is inhibited, thereby facilitating corticostriatal synaptic plasticity (Stipanovich et al., 2008). Carriers of two copies of the T allele (TT carriers) show greater DARPP-32 mRNA expression (Meyer-Lindenberg et al., 2007), which has been linked to a superior ability to learn about action that lead to positive outcomes relative to individuals with at least one C allele (C carriers) (Doll et al., 2011; Frank et al., 2007, 2009). Given the anatomical evidence suggesting that the BG/DA feedback loop would primarily modulate phasic bursts of DA (Joel and Weiner, 2000; Lee et al., 2004), we targeted DARPP-32 as our gene of interest.

We also collected DRD2 and COMT gene labels as part of a suite we commonly collect for the purpose of replication. DRD2 has been associated with D_2 receptor affinity. D_2 receptors are primarily expressed by striatopallidal (indirect NoGo) medium spiny neurons, which are sensitive to fluctuations of DA below baseline levels (Hirvonen et al., 2009). Thus, genetic variation of DRD2 has been argued to modulate learning from negative RPEs, coded as phasic dips in DA below baseline levels (Doll et al., 2011; Frank et al., 2007, 2009; Shen and Flajolet, 2008). The COMT gene codes for an enzyme that breaks down extracellular DA in prefrontal cortex (Meyer-Lindenberg et al., 2005, 2007). Thus, genetic variation in COMT has been argued to modulate individual differences in DA levels, and D_1 receptor availability in prefrontal cortex, which has been shown to drive individual differences in working memory function (Doll et al., 2011; Frank et al., 2007, 2009).

Training phase performance

Participants experienced an average of 5 training blocks (mean 4.7, sem 0.1), with no effect of DARPP-32 gene group on the number of blocks (p>0.5). We assessed performance during the training phase by entering training block (first, last), stimulus pair ($A_{fc}B_{fc}$, $C_{fc}D_{fc}$, $E_{fc}F_{fc}$) and DARPP-32 gene group (C, TT) as factors in a multilevel logistic regression. The omnibus ANOVA revealed a main effect of stimulus pair indicating differential performance among training pairs, a main effect of training block indicating differential performance between first and last blocks, and a stimulus pair by DARPP-32 gene group interaction indicating DARPP-32 gene groups did not perform equally across all stimulus pairings (main effect of stimulus pair: $\chi^2(2) = 29.5$, p<0.01; main effect of training block: $\chi^2(1) = 86.9$, p<0.01; stimulus pair by DARPP-32 gene group interaction: $\chi^2(2) = 16.5$, p<0.01, all other effects n.s). Follow up contrasts show that accuracy improved from first to last block of training (Beta=0.51, p<0.01). In keeping with the stochastic nature of the task, performance was better on $A_{fc}B_{fc}$ than $C_{fc}D_{fc}$ trials (Beta=0.40, p<0.01), and performance was worse on $E_{fc}F_{fc}$ than $C_{fc}D_{fc}$ trials (Beta=-0.40, p<0.01). The DARPP-32 TT group was significantly better on $A_{fc}B_{fc}$ trials (Beta=0.30, p<0.01), but gene group performance was not differentiable otherwise. As such, we included $A_{fc}B_{fc}$ training performance as a covariate when analyzing test performance to control for differential training experience between DARPP-32 gene groups.Genetic predictors of reinforcement learning

Previous work has linked DARPP-32 and DRD2 genotypes to individual differences in learning from positive and negative outcomes (Doll et al., 2011; Frank et al., 2007, 2009). We investigate the relationship between learning and genotype by assessing performance on the most difficult test phase trials, where both options had expected values that were either positive ($A_{fc}C_{fc}$, $A_{fc}E_{fc}$, $C_{fc}E_{fc}$), or negative ($B_{fc}D_{fc}$, $B_{fc}F_{fc}$, $D_{fc}F_{fc}$), by entering DARPP-32 gene group (C, TT), DRD2 gene group (CC, T), and stimulus pair valence (positive, negative) as factors in a multilevel logistic regression (see **Error! Reference source not found.**).

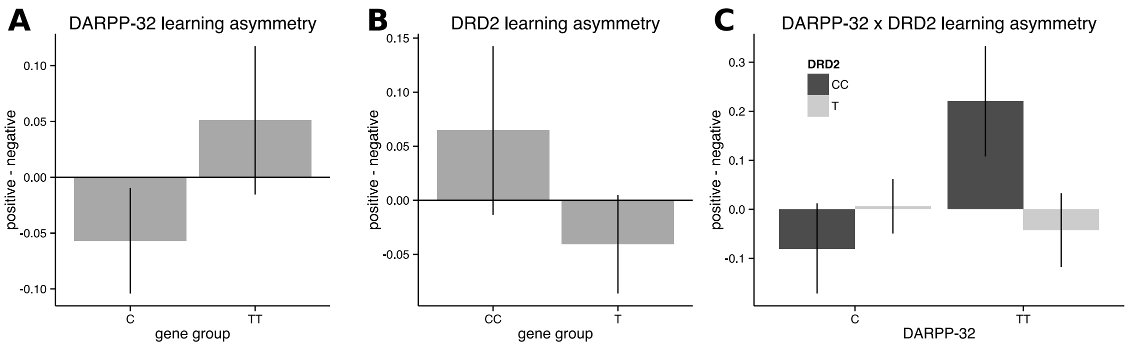

Figure S1 related to Figures 1C and 4C: Individual differences are predicted by DARPP-32 and DRD2. The y-axis represent the mean residual variance in the difference between performance on positive and negative high conflict trials after accounting for relevant terms from the hierarchical logistic regression. Thus, positive values indicate better performance on rewarding relative to non-rewarding high conflict trials. Error bars represent standard error of the mean residual variance. **(A)** Residual variance after accounting for variance associated with DRD2. DARPP-32 TT carriers show better discrimination among rewarding options, whereas C carriers show better discrimination among non-rewarding options. **(B)** Residual variance after accounting for variance associated with DARPP-32. DRD2 T carriers show better discrimination among non-rewarding options, whereas CC carriers show better discrimination among rewarding options. **(C)** Residual variance after accounting for variance associated with the DARPP-32 by DRD2 interaction. Carriers of both DARPP-32 TT and DRD2 CC genotypes are disproportionately better at discriminating among rewarding options.

This analysis revealed a significant DARPP-32 by valence interaction after DRD2 effects were accounted for, with DARPP-32 TT carriers exhibiting better discrimination among positive options, and C-carriers exhibiting better discrimination among negative options ($\chi^2(1) = 6.5$, p=0.01, **Error! Reference source not found.**A). There was also a significant DRD2 by valence interaction after DARPP-32 effects were accounted for, with DRD2 T carriers exhibiting better discrimination among negative options, and DRD2 CC carriers showing better discrimination among positive options ($\chi^2(1) = 4.6$, p=0.03, **Error! Reference source not found.**B). There was also a significant three-way DARPP-32 by DRD2 by valence interaction, indicating that individuals carrying both the DARPP-32 TT and DRD2 CC alleles were disproportionately better at discriminating among positive stimuli ($\chi^2(1) = 4.4$, p=0.04, **Error! Reference source not found.**C). These results replicate previous work that linked genes associated with dopaminergic striatal plasticity to individual differences in learning from either positive or negative outcomes (Doll et al., 2011; Frank et al., 2007, 2009). Importantly, these results demonstrate that DARPP-32 genotype predicts differential learning in line with the learning rate parameter manipulations applied to the computational model.

Behavior is consistent with derived value structure

Figure3A of the main text depicts the value structure derived from the behavioral choice bias pattern. There, no-choice values take on the true expected value of each option (e.g. $nc_{80\%} = E[A_{nc}]$). Free-choice values accommodate the effects of choice by taking the true expected value of each option adjusted according to choice biases for each option (e.g. $fc_{80\%} = E[A_{fc}] + b_A$). Assuming that the difference between option values determines the reliability with which the better option will be selected, which is captured by the Softmax action selection mechanism, this value structure predicts not only the observed choice bias preferences, but also the degree to which each option should be preferred over any of the others.

As noted in the main text, the discrepancy between equally rewarded options (e.g. $b_{80\%} = fc_{80\%} - nc_{80\%}$, or $b_{20\%} = fc_{20\%} - nc_{20\%}$) implies a constant value discrepancy among trials where those options are paired with the same alternative (e.g. $fc_{80\%} - fc_{30\%} = (nc_{80\%} + b_{80\%}) - fc_{30\%}$, and $fc_{20\%} - nc_{60\%} = (nc_{20\%} + b_{20\%}) - nc_{60\%}$). We probed for this predicted pattern by assessing accuracy on trials involving either the most or the least rewarding free-choice and no-choice options. We begin by focusing on the most-rewarding options, entering root option $(A_{fc},\ A_{nc})$, and paired option $(C_{fc}, E_{fc}, \dots\ D_{nc})$ as factors in a multilevel logistic regression (see Figure 3B: A_{fc}, and A_{nc}). As noted in the main text, this analysis revealed a main effect of root option, indicating an A_{fc} performance benefit; however, there was no evidence of a root by paired option interaction, indicating that A_{fc} performance gains were consistent across all paired options (main effect of root option: $\chi^2(1) = 29.23$, p<0.01; main effect of paired option: $\chi^2(7) = 138.02$, p<0.01; root by paired option interaction: $\chi^2(7) = 9.25$, p>0.2). Furthermore, adjusting A_{fc} trial accuracy by the behaviorally quantified choice bias (Figure 3B: $A_{fc} - b_A$) rendered performance indistinguishable from A_{nc} trials, indicating that A_{fc} performance benefits across all options pairings were consistent with the choice bias (main effect of root: $\chi^2(1) = 0.15$, p>0.6; main effect of pairing: $\chi^2(7) = 127.43$, p<0.01; root by pairing interaction: $\chi^2(7) = 9.26$, p>0.2).

The predicted preference pattern was also observed across trials that included the least rewarding options. We entered root option (B_{fc}, B_{nc}), and paired option $(C_{fc}, E_{fc}, \dots\ D_{nc})$ as factors in a multilevel logistic regression (see Figure 3C: B_{fc}, and B_{nc}). As predicted by the insignificant b_B choice bias, we observe a non-significant root by paired option interaction (main effect of root option: $\chi^2(1) = 0.98$, p>0.3; main effect of paired option: $\chi^2(7) = 154.03$, p<0.01; root by paired option interaction: $\chi^2(7) = 2.32$, p>0.9). Again, adjusting B_{fc} trial accuracy by the behaviorally quantified choice bias (Figure 3C: $B_{fc} - b_B$) rendered performance indistinguishable from B_{nc} trials, indicating that any B_{fc} performance discrepancies are consistent with the choice bias across all option pairings (main effect of root option: $\chi^2(1) = 0.69$, p>0.4; main effect of paired option: $\chi^2(7) = 150.70$, p<0.01; root by paired option interaction: $\chi^2(7) = 1.69$, p>0.9).

A second prediction made by the derived value structure comes from the discrepancy among rewarding and non-rewarding options. This discrepancy is greater for free-choice options than it is for no-choice options, owing to the amplified value of rewarding free-choice options (e.g. $fc_{80\%} - fc_{30\%} > nc_{80\%} - nc_{30\%}$). However, the discrepancy among options of equal valence is roughly equivalent across choice conditions ($fc_{80\%} - fc_{70\%} \approx nc_{80\%} - nc_{70\%}$). Thus, the derived value structure implies that participants should be more accurate on trials where rewarding and non-rewarding free-choice options are paired. We probed for this predicted pattern by assessing performance on trials involving the most rewarded and least rewarded options and where both options were either free-choice or no-choice.

We first focus on trials involving the most rewarded option, entering root option (A_{fc}, A_{nc}), and the alternative option's valence (rewarding, non-rewarding) as factors in a multilevel logistic regression (see Figure S2A). As predicted, there was a significant root option by valence interaction, indicating that performance improved disproportionately when A_{fc} was paired with a non-rewarding free-choice alternative (main effect of root option: $\chi^2(1) = 4.52$, p=0.03; main effect of valence:$\chi^2(1) = 9.91$, p<0.01; root option by valence interaction: $\chi^2(1) = 4.31$, p=0.04). Furthermore, adjusting A_{fc} performance by the behaviorally quantified choice bias (Figure S2A: $A_{fc} - b_A$), where the bias effect for each paired option was computed as the choice bias discrepancy between both alternatives (i.e., $b_A - b_C, b_A - b_E, b_A - b_F, b_A - b_D$), rendered performance indistinguishable from A_{nc} trials (main effect of root option: $\chi^2(1) = 0.09$, p>0.75; main effect of valence: $\chi^2(1) = 3.05$, p=0.08; root option by valence interaction: $\chi^2(1) = 1.18$, p>0.25).

The predicted preference pattern was also observed across trials that included the least rewarding options. We entered root option (B_{fc}, B_{nc}), and the alternative option's valence (rewarding, non-rewarding) as factors in a multilevel logistic regression (see Figure S2B: B_{fc}, and B_{nc}). As predicted, there was a significant root option by valence interaction, indicating that performance improved disproportionately when B_{fc} was paired with a rewarding free-choice alternative (main effect of root option: $\chi^2(1) = 15.48$, p<0.01; main effect of valence:$\chi^2(1) = 41.10$, p<0.01; root option by valence interaction: $\chi^2(1) = 10.03$, p<0.01). And again, adjusting B_{fc} performance by the behaviorally quantified choice bias (Figure S2B: $B_{fc} - b_B$), where the bias effect for each paired option was computed as the choice bias differences of each stimulus (i.e., $b_C - b_B, b_E - b_B, b_F - b_B, b_D - b_B$), rendered performance indistinguishable from B_{nc} trials (main effect of root option: $\chi^2(1) = 0.64$, p>0.4); main effect of valence: $\chi^2(1) = 5.00$, p=0.03; root option by valence interaction: $\chi^2(1) = 0.08$, p>0.75).

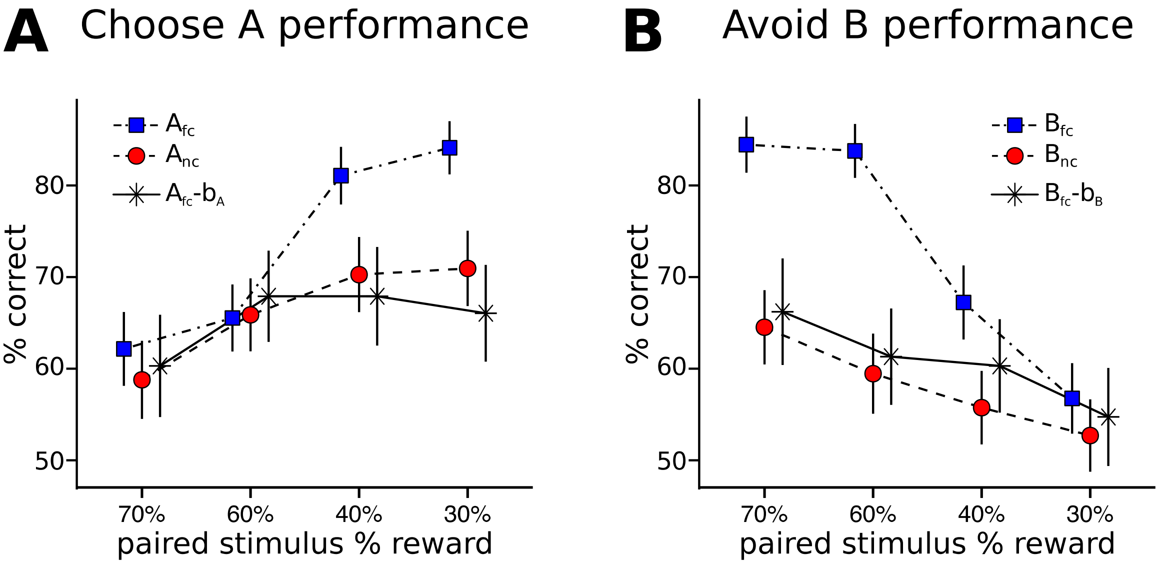

Figure S2 related to Figure 3: Choice bias predicts Choose A and Avoid B performance valence effects. (A) Choose A performance on free-choice (A_{fc}: $A_{fc}C_{fc}$, $A_{fc}E_{fc}$, $A_{fc}F_{fc}$, $A_{fc}D_{fc}$) and no-choice (A_{nc}: $A_{nc}C_{nc}$, $A_{nc}E_{nc}$, $A_{nc}F_{nc}$, $A_{nc}D_{nc}$) trials, and free-choice performance adjusted according to the behavioraly quantified bias ($A_{fc} - b_A$: $A_{fc}C_{fc} - (b_A - b_c)$, $A_{fc}E_{fc} - (b_A - b_E)$, $A_{fc}F_{fc} - (b_A - b_F)$, $A_{fc}D_{fc} - (b_A - b_D)$,). **(B)** Avoid B performance on free-choice (B_{fc}: $C_{fc}B_{fc}$, $E_{fc}B_{fc}$, $F_{fc}B_{fc}$, $D_{fc}B_{fc}$) and no-choice (B_{nc}: $C_{nc}B_{nc}$, $E_{nc}B_{nc}$, $F_{nc}B_{nc}$, $D_{nc}B_{nc}$) trials, and free-choice performance adjusted according to the behavioraly quantified bias ($B_{fc} - b_B$: $C_{fc}B_{fc} - (b_C - b_B)$, $E_{fc}B_{fc} - (b_E - b_B)$, $F_{fc}B_{fc} - (b_F - b_B)$, $D_{fc}B_{fc} - (b_D - b_B)$.

Together, these results demonstrate that participant behavior corresponded with patterns predicted by the choice bias derived value structure in striking detail across a wide range of independent option pairings. These results also show that participants learned the relative values of both free-choice and no-choice options, that preferences were internally consistent across stimulus pairs, and, as predicted by our computational model, that choice bias effects are more pronounced across rewarding options.

An alternative cortico-striatal mechanism

We also consider an alternative mechanism, perhaps via cortico-striatal projections, through which the reported choice bias pattern could emerge whereby striatal activity is shaped via cortical projections such that it uniquely reflects gated actions. Indeed, previous modeling work has applied a similar architecture to the domain of motor action learning (Frank, 2005); however, this architecture does not generalize well to cognitive 'actions' such as the update and maintenance of working memory (Hazy et al., 2006; O'Reilly and Frank, 2006).

Previous work has shown that a gene encoding catechol-O-methyltransferase (COMT) influences DA levels in PFC (Meyer-Lindenberg et al., 2005, 2007), and in turn, PFC-dependent cognitive function such as directed exploration (Frank et al., 2007, 2009) and on-line maintenance of information (Doll et al., 2011). We reasoned that if a cortico-striatal mechanisms was involved in shaping striatal activity, then we should see some effect of COMT on the choice bias.

In contrasts to the strong effects of DARPP-32 on choice bias, we found no effect of COMT on the choice bias when entering COMT gene grouping (Met, ValVal) and choice bias pair $(b_A, b_C, \dots, b_B)$ as factors in a multilevel logistic regression (main effect of COMT: $\chi^2(1) = 0.21$, p=0.64, main effect of pair: $\chi^2(5) = 38.8$, p<0.01, COMT by pair interaction: $\chi^2(5) = 8.3$, p=0.14).

However, replicating previous results (Doll et al., 2011; Frank et al., 2007), we did find that COMT genotype predicted individual differences early in learning. We analyzed inter-trial response behavior across the first five trials of learning for each option pair. Participants could either stay with the same response from the previous trial, or switch to sample the alternate option. This analysis revealed an effect of COMT (Beta=0.18, p=0.03), with ValVal carriers opting to stay with the same response more frequently than Met carriers. Previous work has suggested that Met carriers follow an uncertainty based action selection strategy early in learning, in that they are more likely to sample in accordance with the degree of uncertainty regarding option reward contingencies (Frank et al., 2009). Also in accordance with previous findings, we found no effect of DARPP-32 on inter-trial response behavior (Beta=0.09, p=0.28).

Although the genetic/behavioral data presented here cannot discredit the aforementioned cortico-striatal mechanism with certainty, we did not find any positive evidence to support it; and as such, we favor the choice amplified positive RPE hypothesis put forth in the main text. Notably, this double dissociation between cortical (COMT) and striatal (DARPP-32) DA function suggests that the choice bias pattern emerges primarily as a result of striatal processes.

Supplemental Experimental Procedures

Experimental Design

The Brown University Human Research Committee approved all task procedures. Participants sat in front of a computer screen in a lit room and viewed pairs of visual stimuli that are not easily verbalized (Japanese Hiragana characters). Stimuli were 300x300 pixels, presented in black on a white background.

During the training phase, six different stimulus pairs ($A_{fc}B_{fc}$, $C_{fc}D_{fc}$, $E_{fc}F_{fc}$, $A_{nc}B_{nc}$, $C_{nc}D_{nc}$, $E_{nc}F_{nc}$) were presented in random order, with assignment of Hiragana character to options $A_{fc} - F_{fc}$ and $A_{nc} - F_{nc}$, counterbalanced across subjects. Probabilistic feedback followed option selection. Choosing option A_{fc} lead to positive feedback 80% of the time, whereas choosing option B_{fc} lead to positive feedback only 20% of the time. Options $C_{fc}D_{fc}$ and $E_{fc}F_{fc}$ pairs were less reliable: option C_{fc} was rewarded 70% of the time (D_{fc} was rewarded 30% of the time), and E_{fc} was rewarded 60% of the time (F_{fc} was rewarded 40% of the time). Thus, over the course of training participants should learn to choose options A_{fc}, C_{fc}, and E_{fc}

more often than the paired alternative. Before the training phase of the task began, participants were given the following instructions:

> *Your task is to learn about various symbols. Some symbols will award points more reliably than others, but you'll have to learn which ones those are. On each trial, two symbols will appear on the screen simultaneously. You can select either the symbol on the left using the "S" key, or the symbol on the right using the "K" key. The symbol you select will either award (+1) or lose a point (-1). There's no absolute correct answer, but try to pick symbols that have the best chance of awarding points. At first this might seem difficult, but you'll get lots of practice. On some trials one of the symbols will be selected for you and will be framed in blue. These are called "Match" trials. On "Match" trials, you must select the framed symbol. On other trials you will be free to choose either symbol. These are called "Choose" trials. Regardless of whether you Choose or Match on each trial, your goal is to learn which symbols are more rewarding. Doing so will help you later in the task. Please let the experimenter know if you have any questions or don't fully understand your task. Press the space bar when you're ready to begin.*

On free-choice trials ($A_{fc}B_{fc}$, $C_{fc}D_{fc}$, $E_{fc}F_{fc}$) participants were free to choose either option presented to them. No-choice trials ($A_{nc}B_{nc}$, $C_{nc}D_{nc}$, $E_{nc}F_{nc}$,) were yoked to free-choice trials to ensure identical sampling and reinforcement histories between conditions. The selected option and feedback from each free-choice trial was recorded and used to generate a yoked no-choice trial. For example, if C_{fc} was selected on a $C_{fc}D_{fc}$ trial, and -1 was provided as feedback, a corresponding $C_{nc}D_{nc}$ trial would be generated that forced the selection of C_{nc} (indicated by a blue frame surrounding that option) and provide -1 as feedback. As such, C_{nc} was sampled the same number of times and delivered the same feedback as C_{fc}, and the same follows for the remaining options. Free-choice and no-choice trials were pseudo-randomized within each training block. No-choice trials were presented no later than 5 trials following their yoked free-choice successor to prevent differential learning between conditions.

After each choice, visual feedback text ('+1' in green or '-1' in red) was provided (duration 1 second). Trials were aborted but repeated later if no response was made after three seconds, with 'Too Slow!' feedback text displayed in blue (duration 1 second). Match trials were repeated if the participant did not select the pre-selected stimulus, with 'You must select the framed stimulus' feedback text displayed in black (duration 1 second). Prior to starting the training phase of the experiment, participants were given 6 practice trials of both free-choice and no-choice trials while the experimenter was present to ensure they understood the instructions.

Participants completed a minimum of 4 and maximum of 6 training block, with each block delivering 20 exposures to each of the 6 option pairs, for a total of 120 trials per block. We enforced a performance criterion evaluated at the end of each block to ensure that all participants were at approximately same performance level before advancing to the test phase. Due to differences in the feedback reliability across option pairs, we used different criteria for each pair (65% selection of A_{fc}, 60% selection of C_{fc}, 50% selection of E_{fc}).

Participants could advance to the test phase of the task after completing a minimum of 4 blocks and exceeding the practice criterion, or after 6 blocks (720 trials). Before starting the test phase of the experiment, participants were given the following instructions:

> *Great Job! It's time to test what you've learned. Now you'll be free to choose on every trial, but you'll no longer receive any feedback. If you see new combinations of symbols, choose the symbol that "feels" most likely to award points based on what you've learned. If you're not sure which one to pick, just go with your gut instinct. Press the space bar when you're ready to begin.*

Participants were subsequently tested on a full permutation of all possible option pairings (eight pairings of each choice bias pair, and four repetitions of all other pairings) in random order. Participants were free to choose either option on each test trial, but were no longer provided feedback.

Statistical models

Statistical tests were performed using hierarchical logistic regression models (`lme4` package in R), using trial response accuracy (selection of more rewarding option) as the dependent variable. Independent variables of interest were entered as fixed effects, and where appropriate, within-subject effects were entered as random by subject effects. We included DRD2 gene grouping, race, and $A_{fc}B_{fc}$ training performance as covariates in all analyses that included DARPP-32 gene grouping as a factor to control for potential effects of gene interactions, race, or training experience.

Computational model specification

We employed an extended actor-critic reinforcement learning architecture, which we refer to as Opponent Actor Learning (OpAL), to formally test our hypothesis that free-choice enhances positive RPEs. Biologically inspired neural network models of the BG have demonstrated the importance of considering what can be thought of as an opponent processes between the direct 'Go' and indirect 'NoGo' pathways of the striatum (Frank, 2005; Frank et al., 2004; Hazy et al., 2006). We have distilled the core computations of these models down to a formulaic model specification with a number of free parameters suitable for data driven value estimation.

The classic actor-critic architecture is comprised of a critic that estimates expected values, and an actor that selects actions. The critic's value estimates can be thought of as predictions. Outcomes that turn out better or worse than predicted generate positive and negative RPEs respectively. These RPEs are then used to update the critic's value prediction, and to modify the actor's action weights with aspirations of reliably picking the most appropriate actions in the future.

Like the standard actor-critic framework, OpAL computes the RPE by comparing the critic's expected value with the observed outcome, and uses the RPE as a learning signal to update the critic's future expectation:

$$\delta_t = r_t - V_t \tag{1}$$

$$V_{t+1} \leftarrow V_t + \alpha_c \times \delta_t \tag{2}$$

In a slight departure from conventional notation due to the structure of the task, t represents the current trial. In words, the RPE (δ_t) captures the discrepancy between the predicted reward (V_t) and the observed reward (r_t) for the current trial. The critic's predicted reward estimate is updated proportionally to the RPE according to the critic’s learning rate (α_c).

The OpAL actor extends the standard actor to include both Go (G) and NoGo (N) action weights, which capture the distinctions, and the functional implications thereof, between the direct and indirect pathways respectively. Like the standard actor-critic, action weights are adjusted using the same RPE signal used to update the critic:

$$G_{t+1} \leftarrow G_t + \alpha_g \times [G_t \times +\delta_t] \tag{3}$$

$$N_{t+1} \leftarrow N_t + \alpha_n \times [N_t \times -\delta_t] \tag{4}$$

where α_g and α_n are independent Go and NoGo actor learning rates. The OpAL actor update differs from the standard actor update in two important ways. First, the RPE signal (δ_t) has opposite effects on G and N action weights. This is intended to mirror dopamine's differential effects on D_1 receptor expressing Go cells in the direct pathway, to which dopamine is generally excitatory, and D_2 receptor expressing NoGo cells in the indirect pathway, to which dopamine is inhibitory. Thus, positive RPEs increase G weights while simultaneously decreasing N weights, and vice versa for negative RPEs. Second, action weights are updated in proportion to not only the RPE, but also with respect to the action weight itself (G_t or N_t). This incorporates the notion that, in addition to dopaminergic input, activity at the cell itself governs the rate of synaptic change that can occur. Including both the action weight and RPE terms in the update rule captures the notion of three-factor Hebbian learning, were synaptic change depends on presynaptic activation, postsynaptic activation, and dopamine (Reynolds et al., 2001).

The experimental paradigm presented here always involved a choice between two options. We compute the probability of choosing option a_1, where $p(a_2) = 1 - p(a_1)$, according to the Softmax action selection rule:

$$Q(a_1) = \beta_g \times G(a_1) - \beta_n \times N(a_1) \tag{5}$$

$$p(a_1) = \frac{1}{1 + e^{Q(a_2) - Q(a_1)}} \tag{6}$$

Note that $p(a_1)$ depends on a linear combination of $G(a_1)$ and $N(a_1)$ weights, where Go and NoGo weights for a_1 are scaled according to independent β_g and β_n parameters respectively.

Generative computational model

To explore the consequences of amplifying positive free-choice RPEs we added a single parameter to the core OpAL model (α_{fc+}) that modulated positive free-choice RPEs (see **Error! Reference source not found.** for model parameters used to generate data). We then exposed the model to the experimental task, allowing the model to generate its own response on each trial and learn accordingly. The following modified update rules were applied to the actor on free-choice trials where $\delta_t > 0$ (using equations (3) and (4) otherwise):

$$G_t \leftarrow G_t + [\alpha_{fc+} \times \alpha_g] \cdot [G_t \times +\delta_t] \tag{7}$$

$$N_t \leftarrow N_t + [\alpha_{fc+} \times \alpha_n] \cdot [N_t \times -\delta_t] \tag{8}$$

Figure	$\boldsymbol{\beta_g}$	$\boldsymbol{\beta_n}$	$\boldsymbol{\alpha_c}$	$\boldsymbol{\alpha_g}$	$\boldsymbol{\alpha_n}$	$\boldsymbol{\alpha_{fc+}}$
Figure 2B Figure 4A	1	3	0.05	0.15	0.15	as noted
Figure 4B $\alpha_g > \alpha_n$	1	3	0.05	0.11	0.025	1.22
Figure 4B $\alpha_g < \alpha_n$	1	3	0.05	0.025	0.15	1.22
Figure S4	1	3	0.05	[0.01-0.2]	[0.01-0.2]	as noted

Table S1 related to Figures 2 & 4: Model parameters used to generate figures. Softmax action selection parameters for the Go and NoGo weights (β_g and β_n), critic learning rate (α_c), Go and NoGo weight learning rates (α_g and α_n), and positive choice RPE modulation parameter (α_{fc+}).

As noted in the main text, as α_{fc+} increased, and learning was balanced across Go and NoGo weights ($\alpha_g = \alpha_n$), the choice bias increased across rewarding stimuli (see Figure 2A). However, when $\alpha_g \neq \alpha_n$, the choice bias pattern across rewarding stimuli shifted dramatically. Figure 4B of the main text depicts the choice bias for individual $\alpha_g > \alpha_n$ and $\alpha_g < \alpha_n$ values; however, the pattern depicted there holds more generally. Each point in Figure S3 represents the regression slope of the model generated choice bias across positive options as a function of the $\alpha_g - \alpha_n$

asymmetry. This analysis demonstrates a strong relationship between the $\alpha_g - \alpha_n$ asymmetry and choice bias (r(88)=-0.8, p<0.01); when $\alpha_g < \alpha_n$ the choice bias slope is positive ($b_A < b_C < b_E$), and when $\alpha_g > \alpha_n$ the choice bias slope is negative ($b_A > b_C > b_E$).

Model comparison and parameter estimation

Although the choice bias patterns generated by extending the base OpAL model to include the α_{fc+} parameter mirrored behavioral data, we also considered alternative mechanisms formalized by additional variants of the base OpAL model. Each model's free parameters were optimized using the `nlminb` function in R to search for a parameter set that minimized the discrepancy between the empirical data and the model's predicted response on each trial, a processes that was repeated multiple times with random starting points to avoid local minimal. The search space for each parameter was unbounded ($[-\infty, +\infty]$) so as to recover a normally distributed parameter set. Each model mapped the optimizer's parameter proposals onto a suitable range: Softmax β weights were mapped onto a non-negative range ($\beta = e^{p_\beta}$, where p_β is the parameter proposed by the optimizer), and learning rates were constrained to range between 0 and 1 ($\alpha = \frac{1}{1+e^{-p_\alpha}}$ where p_α is the learning rate parameter proposed by the optimizer).

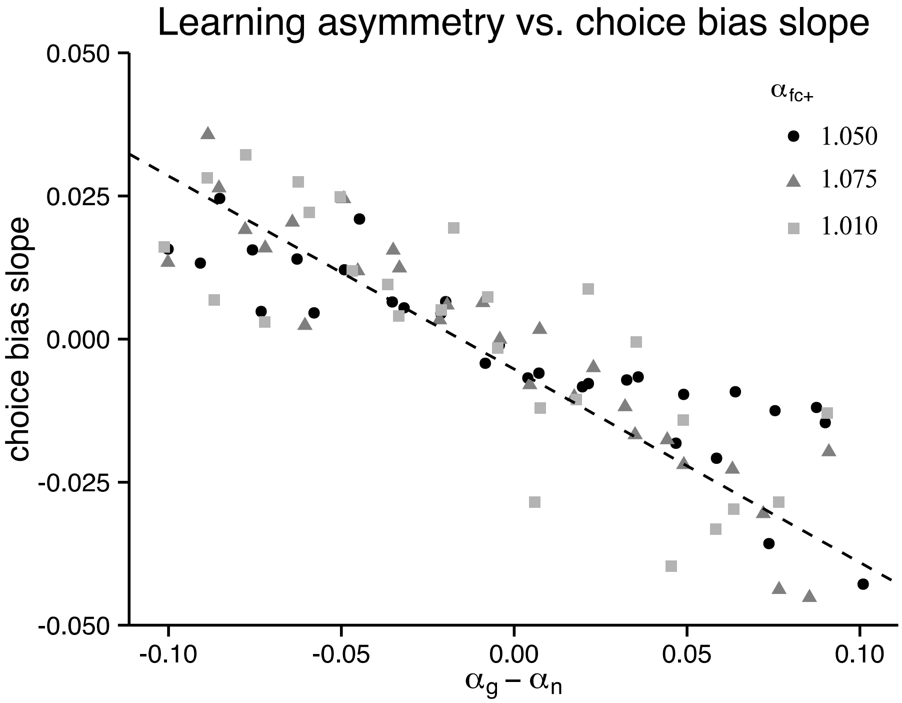

Figure S3 related to Figure 4B: Choice bias as a function of $\alpha_g - \alpha_n$ learning rate asymmetry. Each point represents a dataset generated by a different parameter set supplied to the OpAL + α_{fc+} model. Points vary along the x-axis according to $\alpha_g - \alpha_n$, and vary along the y-axis according to the slope of a regression on rewarding option choice biases.

Model fit was assessed using AIC values, calculated for each participant's data as:

$$AIC_{m.i} = 2k_m + 2 \cdot L_{m.i}(\theta|D_i)] \tag{9}$$

where k_m is the number of free parameters in model m, and $L_{m.i}(\theta|D_i)$ is the negative log-likelihood of the parameter set θ given subject i's data, D_i. Thus, lower $AIC_{m.i}$ values indicate a better fit for model m given participant i's data. Model fits were compared using Bayesian model comparison (`spm_BMS` function in SPM8)(Stephan et al., 2009), and AIC weights (Wagenmakers and Farrell, 2004).

The core OpAL model served as the root of our model comparison. This model includes Go and NoGo Softmax parameters (β_g and β_n), Go and NoGo actor learning rate parameters (α_g and α_n), and a critic learning rate parameter (α_c). We note that traditional actor-critic and Q-learning models were also explored; however, none of those models were capable of capturing the behavioral choice bias pattern. As such, we focus solely on variants of the OpAL model for clarity and because of the model's biologically motivated structure.

Our hypothesis that positive free-choice RPEs are amplified was most simply formalized by extending the OpAL model with a single modulatory learning rate parameter α_{fc+}. This parameter was incorporated into the value update functions for both Go and NoGo weights:

$$\alpha_{fc.g} = 1/\left(1 + e^{-(p_g + p_{\alpha_{fc+}})}\right) \tag{10}$$

$$\alpha_{fc.n} = 1/\left(1 + e^{-(p_n + p_{\alpha_{fc+}})}\right) \tag{11}$$

where p_g, p_n, and p_{fc+} are the Go, NoGo, and free-choice learning rate modulation parameters proposed by the optimizer, and $\alpha_{fc.g+}$ and $\alpha_{fc.n+}$ are the effective parameters used by the model on free-choice trials with $\delta_t > 0$. Incorporating α_{fc+} into the model by adding it to p_g and p_n prior to sigmoidal transformation ensured that the modulated learning rates were bound between 0 and 1.

We also considered the possibility that participants were simply more engaged free-choice trials. We formalized this by adding a single parameter to the base OpAL model, α_{fc}, which modulated both the α_g and α_n learning rates on all free-choice trials, irrespective of the RPE's sign.

We also considered the possibility that choice modulated both positive and negative RPEs independently. We formalized this by extending the OpAL model to include both α_{fc+} and α_{fc-} free-choice learning rate modulatory parameters, which were applied to positive and negative free-choice RPEs respectively.

Finally, we considered the possibility that the choice bias may not be an effect of learning at all, but may emerge from an action selection mechanism. To investigate this, we extended the OpAL model to include independent Softmax parameters for both free-choice ($\beta_{fc.g}$, $\beta_{fc.n}$) and no-choice options ($\beta_{nc.g}$, $\beta_{nc.n}$).

Bayesian model selection strongly favored the model with both α_{fc+} and α_{fc-} parameters, as did the AIC weights (see Table S2), suggesting that group behavior was best explained by modulating learning for both positive and negative free-choice RPEs. An analysis of this model's parameter estimates revealed that α_{fc+} was significantly greater than α_{fc-} (paired t(73)=3.32, p=0.001, C.I=[1.8,7.0]), and that the majority of subjects were best fit by an $\alpha_{fc+} > \alpha_{fc-}$ asymmetry (57 of 74 participants, binomial test p<0.001). Additional analyses show that $\alpha_{fc+} > 0$ (t(73)=2.29, p=0.025, C.I=[0.67,9.68]), indicating data was best fit by amplifying the learning rate when positive free-choice RPEs were encountered. Conversely, α_{fc-} could not be distinguished from zero (t(73)=0.33, p=0.74, C.I=[-3.9,5.4]). In sum, optimized parameter estimates are in accordance with our hypothesis that positive free-choice RPEs are preferentially amplified.

A detailed inspection of non-rewarding choice biases shows some degree of inter-participant variance, with a small subset of study participants exhibiting a choice bias for those options (N=16 out of 80 total). Although most participants were indifferent on choice bias trials involving negative options, some participants showed a strong preference for free-choice or no-choice options. However, these biases were unsystematic, with participants showing a bias for only a single negative option, or exhibiting a free-choice preference for one negative option and a no-choice preference for another. The α_{fc-} parameter allowed some of this variance to be accounted for, resulting in a better fit for the model that included it as a free parameter; but, we could not identify genetic or behavioral predictors of α_{fc-} estimates, suggesting that individual differences in negative option choice biases involved mechanisms beyond the BG. We tentatively suggest that some negative option biases may be driven by a rule-based strategy, not the value based strategy depicted in Figure 3A. Further research will be required to expose the underlying mechanism driving choice biases in these individuals.

We analyzed the best-fit model in terms of DARPP-32 groupings. This revealed a trending effect of DARPP-32 on $\alpha_g - \alpha_n$ learning rate asymmetry (t(71)=1.78, p=0.08, C.I=[-0.02, 0.43]). To further probe these effects, we used a model comparison approach by comparing DARPP-32 group fit on models that were forced to adhere to either a $\alpha_g > \alpha_n$ or a $\alpha_g < \alpha_n$ learning rate asymmetry. As outlined in Table S3, DARPP-32 TT carriers were best fit by the $\alpha_g > \alpha_n$ model, whereas C carriers were best fit by the $\alpha_g < \alpha_n$ model. Together, these results show that the Go/NoGo learning rate asymmetries differ as predicted according to DARPP-32 gene group.

Model	k	AIC	AIC ω	Exceedance Probability
OpAL	5	325.41	1e-9	0
OpAL + α_{fc+}	6	287.41	4.5e-2	0
OpAL + α_{fc}	6	289.54	1.7e-2	0
OpAL + α_{fc+}+ α_{fc-}	7	280.88	0.94	1
OpAL + $\beta_{fc.g}$ + $\beta_{fc.n}$	8	294.71	1.5e-3	0

Table S2 related to Figure 4C: Model Fit. The number of free parameters ($\boldsymbol{k}$), mean AIC value across all subjects (AIC, lower scores indicate better fit), AIC weights when all models are included for comparison (AIC $\boldsymbol{\omega}$, values closer to 1 indicate best fit model), and the exceedance probabilities according to Bayesian model selection (values closer to 1 indicate best model fit) for OpAL model variants. All fit indicators point to OpAL + $\boldsymbol{\alpha_{fc+}}$ + $\boldsymbol{\alpha_{fc-}}$ as the model that explains behavioral data best.

DARPP-32	Model	k	AIC	AIC ω	Exceedance Probability
TT	$\alpha_g > \alpha_n$	7	269.47	0.61	1
	$\alpha_g < \alpha_n$	7	271.79	0.39	0
C	$\alpha_g > \alpha_n$	7	299.09	0.42	0.34
	$\alpha_g < \alpha_n$	7	297.88	0.58	0.66

Table S3 related to Figure 4B & 4C: DARPP-32 gene group mode fit. DARPP-32 gene group fits were compared in isolation from one another, where the OpAL + $\boldsymbol{\alpha_{fc+}}$ + $\boldsymbol{\alpha_{fc-}}$ model was constrained to force either $\boldsymbol{\alpha_g > \alpha_n}$ or $\boldsymbol{\alpha_g < \alpha_n}$ asymmetries. The number of free parameters ($\boldsymbol{k}$), mean AIC value across all subjects (AIC, lower scores indicate better fit), AIC weights when both models are included for comparison (AIC $\boldsymbol{\omega}$, values closer to 1 indicate best fit model), and the exceedance probabilities according to Bayesian model selection (values closer to 1 indicate best model fit) for OpAL model variants. All fit indicators show that DARPP-32 TT carriers are best fit by an $\boldsymbol{\alpha_g > \alpha_n}$ asymmetry, whereas C carriers are best fit by an $\boldsymbol{\alpha_g < \alpha_n}$ asymmetry.

Supplemental References

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.

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

Frank, 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.

Frank, M.J., Moustafa, A. a, Haughey, H.M., Curran, T., and Hutchison, K.E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. U. S. A. *104*, 16311–16316.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. *12*, 1062–1068.

Hazy, T.E., Frank, M.J., and O'Reilly, R.C. (2006). Banishing the homunculus: making working memory work. Neuroscience *139*, 105–118.

Hirvonen, M.M., Laakso, A., Någren, K., Rinne, J.O., Pohjalainen, T., and Hietala, J. (2009). C957T polymorphism of dopamine D2 receptor gene affects striatal DRD2 in vivo availability by changing the receptor affinity. Synapse *63*, 907–912.

Joel, D., and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience *96*, 451–474.

Lee, C.R., Abercrombie, E.D., and Tepper, J.M. (2004). Pallidal control of substantia nigra dopaminergic neuron firing pattern and its relation to extracellular neostriatal dopamine levels. Neuroscience *129*, 481–489.

Meyer-Lindenberg, A., Kohn, P.D., Kolachana, B., Kippenhan, S., McInerney-Leo, A., Nussbaum, R., Weinberger, D.R., and Berman, K.F. (2005). Midbrain dopamine and prefrontal function in humans: interaction and modulation by COMT genotype. Nat. Neurosci. *8*, 594–596.

Meyer-Lindenberg, A., Straub, R.E., Lipska, B.K., Verchinski, B.A., Goldberg, T., Callicott, J.H., Egan, M.F., Huffaker, S.S., Mattay, V.S., Kolachana, B., et al. (2007). Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. J. Clin. Invest. *117*, 672–682.

O'Reilly, R.C., and Frank, M.J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. *18*, 283–328.

Reynolds, J.N.J., Hyland, B.I., and Wickens, J.R. (2001). A cellular mechanism of reward-related learning. Nature *413*, 67–70.

Shen, W., and Flajolet, M. (2008). Dichotomous dopaminergic Control of Striatal Synaptic Plasticity. Sci. … 848–851.

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.

Stipanovich, A., Valjent, E., Matamales, M., Nishi, A., Ahn, J.-H.H., Maroteaux, M., Bertran-Gonzalez, J., Brami-Cherrier, K., Enslen, H., Corbillé, A.-G.G., et al. (2008). A phosphatase cascade by which rewarding stimuli control nucleosomal response. Nature *453*, 879–884.

Wagenmakers, E.-J., and Farrell, S. (2004). AIC model selection using Akaike weights. Psychon. Bull. Rev. *11*, 192–196.