Behavioral/Cognitive

Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning

Bradley B. Doll,[1,2] **Kevin G. Bath,**[3] **Nathaniel D. Daw,**[4,5*] **and Michael J. Frank**[3,6*]

[1]Center for Neural Science, New York University, New York, New York 10003, [2]Department of Psychology, Columbia University, New York, New York 10027, [3]Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, Rhode Island 02912, [4]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08540, [5]Department of Psychology, Princeton University, Princeton, New Jersey 08540, and [6]Brown Institute for Brain Science, Brown University, Providence, Rhode Island 02912

Considerable evidence suggests that multiple learning systems can drive behavior. Choice can proceed reflexively from previous actions and their associated outcomes, as captured by "model-free" learning algorithms, or flexibly from prospective consideration of outcomes that might occur, as captured by "model-based" learning algorithms. However, differential contributions of dopamine to these systems are poorly understood. Dopamine is widely thought to support model-free learning by modulating plasticity in striatum. Model-based learning may also be affected by these striatal effects, or by other dopaminergic effects elsewhere, notably on prefrontal working memory function. Indeed, prominent demonstrations linking striatal dopamine to putatively model-free learning did not rule out model-based effects, whereas other studies have reported dopaminergic modulation of verifiably model-based learning, but without distinguishing a prefrontal versus striatal locus. To clarify the relationships between dopamine, neural systems, and learning strategies, we combine a genetic association approach in humans with two well-studied reinforcement learning tasks: one isolating model-based from model-free behavior and the other sensitive to key aspects of striatal plasticity. Prefrontal function was indexed by a polymorphism in the *COMT* gene, differences of which reflect dopamine levels in the prefrontal cortex. This polymorphism has been associated with differences in prefrontal activity and working memory. Striatal function was indexed by a gene coding for *DARPP-32*, which is densely expressed in the striatum where it is necessary for synaptic plasticity. We found evidence for our hypothesis that variations in prefrontal dopamine relate to model-based learning, whereas variations in striatal dopamine function relate to model-free learning.

Key words: decision-making; dopamine; genetics; reinforcement learning

Significance Statement

Decisions can stem reflexively from their previously associated outcomes or flexibly from deliberative consideration of potential choice outcomes. Research implicates a dopamine-dependent striatal learning mechanism in the former type of choice. Although recent work has indicated that dopamine is also involved in flexible, goal-directed decision-making, it remains unclear whether it also contributes via striatum or via the dopamine-dependent working memory function of prefrontal cortex. We examined genetic indices of dopamine function in these regions and their relation to the two choice strategies. We found that striatal dopamine function related most clearly to the reflexive strategy, as previously shown, and that prefrontal dopamine related most clearly to the flexible strategy. These findings suggest that dissociable brain regions support dissociable choice strategies.

Introduction

Decisions can follow from distinct learning strategies. Choices may arise from previous reinforcement (Thorndike, 1898), reflecting learned values from the past, or from prospective consideration of outcomes potentially obtained in the future (Tolman, 1948), based on current goals. These strategies are computationally described by "model-free" and "model-based" reinforcement learning (RL), the former reflecting habitual behaviors and the latter reflecting goal directed ones (Daw et al., 2005). These distinct computations appear to have partly dissociable neural

Received May 16, 2015; revised Dec. 14, 2015; accepted Dec. 14, 2015.

Author contributions: B.B.D., N.D.D., and M.J.F. designed research; B.B.D. performed research; K.G.B. contributed unpublished reagents/analytic tools; B.B.D. analyzed data; B.B.D., K.G.B., N.D.D., and M.J.F. wrote the paper.

This work was supported by National Institute of Mental Health R01 MH080066-01, National Science Foundation 1460604, and National Institute of Neurological Disorders and Stroke R01NS078784. The authors thank Christina Figueroa, Sean Masters, and Julie Helmers for assistance with data collection.

The authors declare no competing financial interests.

*N.D.D. and M.J.F. contributed equally to this work (order determined by coin flip).

Correspondence should be addressed to Dr. Bradley B. Doll, Center for Neural Science, 4 Washington PL, New York University, New York, NY 10003. E-mail: bradley.doll@nyu.edu.

DOI:10.1523/JNEUROSCI.1901-15.2016

0270-6474/16/361211-12$15.00/0

substrates (Doll et al., 2015a), with dopamine (DA) playing a role in each (Wunderlich et al., 2012; Steinberg et al., 2013; Deserno et al., 2015). While much research supports the hypothesized role of DA in model-free learning, DA contributions to model-based learning are poorly understood.

Model-free learning is thought to follow from prediction error signals of midbrain DA neurons (Glimcher, 2011). By this view, DA signals the discrepancy between reward received and reward expected, modulating the plasticity of striatal targets such that actions leading to good outcomes are reinforced, whereas actions leading to bad ones are punished. Optogenetic manipulations show that instrumental conditioning causally depends on these dopaminergic prediction errors (Steinberg et al., 2013), with opposing effects on striatal cells expressing D1 and D2 receptors driving opposing effects on behavioral approach and avoidance learning and choice (Kravitz et al., 2012; Collins and Frank, 2014).

However, most research detailing how these mechanisms manifest behaviorally has used tasks that do not distinguish model-free from model-based learning. This leaves open the possibility that these mechanisms might contribute via model-based RL (Doll et al., 2012). Indeed, recent work shows that DA plays a role in verifiably model-based learning (Wunderlich et al., 2012; Deserno et al., 2015; Sharp et al., 2015). But it remains unclear what model-based computations DA might impact, and where in the brain these effects lie.

Prefrontal cortex (PFC), and DA modulation therein, is a promising candidate for model-based computation (Daw et al., 2005). DA in PFC is thought to support the recurrent cellular activation underlying working memory (Durstewitz and Seamans, 2008), which may afford the laborious computations of the model-based approach (Daw et al., 2005). Accordingly, disruption of PFC by transcranial magnetic stimulation impairs model-based behavior (Smittenaar et al., 2013). Putative disruption of PFC working memory resources by cognitive load or stress produces similar impairments (Otto et al., 2013a, b). Greater baseline reserves of working memory protect against the latter stress effect (Otto et al., 2013b).

To investigate whether PFC and striatal DA play specific roles in model-based versus model-free learning, we examined behavioral associations with genes involved in DA function in dissociable brain regions. We focus on single nucleotide polymorphisms (SNPs) of two genes implicated in DA functions in prefrontal cortex (*COMT*), and striatum (*DARPP-32*), both of which have been implicated in distinct aspects of RL, such as working memory versus incremental learning (Frank et al., 2007; Doll et al., 2011; Collins and Frank, 2012; Cockburn et al., 2014), but not in a context that formally distinguished model-based from model-free learning. We used a task that combines features from two widely studied learning tasks: a sequential choice task (Daw et al., 2011), which formally dissociates model-based from model-free learning; and a probabilistic selection task (Frank et al., 2004), which probes a key feature of the striatal prediction error mechanism, which is normally (but heretofore not verifiably) viewed as key to model-free RL. This latter task has succeeded in capturing differences in learning from rewards and nonrewards and relating this learning to striatal DA effects on cells expressing D1 and D2 receptors (Frank et al., 2004, 2007), consistent with the view that DA reward prediction error signals drive model-free learning (Frank, 2005; Collins and Frank, 2014). These features may provide a more nuanced measure of striatal learning, helping to clarify the regional specificity of dopaminergic effects on sequential decision making, and the nature of the learning probed by each of these widely studied tasks.

Materials and Methods

Subjects. A total of 171 healthy subjects (98 females, mean age = 22.6 years, SD = 4.7 years) recruited from the Brown University and Providence, Rhode Island community, completed the experiment and were paid $10. Transfer phase data were lost from one subject due to a computer error.

Genes. Of the 171 subjects who participated in the study, 105 self-identified as Caucasian (58 females, mean age = 22.8 years, SD = 4.6 years). All reported analyses control for racial group. Here we report the frequencies of the alleles in the group as a whole, and in the Caucasian subset.

COMT (rs4680) Val/Val, Val/Met, Met/Met: 56, 80, 33 (Caucasian subset: 31, 49, 24). Genotyping failed for 2 subjects. *DARPP-32* (rs907094) C/C, C/T, T/T: 27, 71, 68 (Caucasian subset: 7, 40, 55). Genotyping failed for 5 subjects.

The distribution of alleles in neither SNP deviated from Hardy-Weinberg equilibrium (*COMT*: $\chi^2 = 0.21$, $p = 0.65$, Caucasian subset: $\chi^2 = 0.3$, $p = 0.58$; *DARPP-32*: $\chi^2 = 1.3$, $p = 0.25$, Caucasian subset: $\chi^2 = 0.01$, $p = 0.92$).

Across the entire sample, *COMT* Met alleles and *DARPP-32* T alleles were significantly correlated (Spearman $\rho = 0.19$, $p = 0.015$), although this relationship was not reliable in the subset of Caucasian subjects ($\rho = 0.15$, $p = 0.13$). All analyses control for this correlation by assessing cognitive effects of both SNPs in the same statistical models (partialling out any shared variance). We control for potential population stratification effects by including race as a covariate in regression analyses of behavior and of RL model parameters.

DNA collection, extraction, and genotypic analysis. Genomic DNA was collected using Oragene saliva collection kits (DNA Genotek) and purified using the manufacturer's protocol. For genotyping, we used TaqMan 5′ nuclease SNP assays (ABI) for the rs907094 (DARPP32) and rs4680 (*COMT*) SNPs. Assays were performed in duplicate on an CFX384 apparatus (Bio-Rad) in real-time PCR mode using standardized cycling parameters for ABI Assays on Demand. Fluorescence was then analyzed using the allelic discrimination function in the CFX software. Amplification curves were visually inspected for each of the assays that led to determination of the genotype. All samples were required to give clear and concordant results, and all samples that did not were rerun and/or reextracted until they provided clear genotype calls.

Behavioral task. Subjects completed 300 trials of a two-step sequential decision task followed by a transfer phase, preceded by task instructions and 20 practice trials with unique stimuli. The sequential learning task (Daw et al., 2011) measures model-based relative to model-free control: computational characterizations of two of the brain's multiple decision systems. This characterization captures the distinction between goal-directed and habitual behavior (Daw et al., 2005; Doll et al., 2012; Dolan and Dayan, 2013) measured in classic instrumental conditioning studies using revaluation and latent learning paradigms (for a variant of this task measuring latent learning, see Gläscher et al., 2010).

In the first step of the sequential task, subjects chose between two stimuli (2 s response window). This choice stochastically determined a second set of choices with fixed transition probabilities (0.7 and 0.3; Fig. 1*A*; 2 s response window). Choice at the second stage was followed by reward with a slowly and randomly drifting probability (with reflecting boundaries of 0.25 and 0.75; Fig. 1*B*) in the first 150 trials. One of four sets of reward probability drifts was randomly assigned to each subject (drift assignment did not differ by genotype, $p > 0.3$). Ultimately, reward probabilities drifted to final values that were fixed in the second 150 trials (70% vs 30% in one state, 60% vs 40% in the other). This design feature permitted subjects to learn the values of these stimuli incrementally (ostensibly via model-free updating). We fixed the final values so that we could assess subjects' ability to discriminate between these differential learned reward probabilities in a subsequent transfer phase: models and data suggest that the differential ability to choose the most rewarding actions (in this case, 70%) over those that are more neutral compared with avoidance of the least rewarding actions (30%) depends on striatal

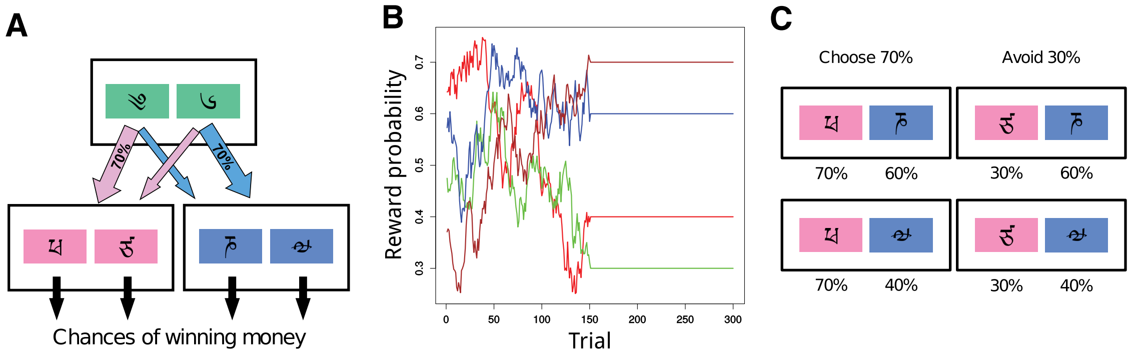

Figure 1. ***A***, Two-step sequential learning task structure. Each of the 300 trials starts with a selection between two first-stage options (green boxes), which produced a set of second-stage options (pink or blue boxes). First-stage options predominantly lead to one set of second-stage options (70% common transitions) but sometimes lead to the other set (30% rare transitions). ***B***, Second-stage choices are rewarded with a randomly diffusing probability for each option. Diffusion ends at trial 150, and probabilities remain fixed for the remainder of trials (70%/30% for one state, 60%/40% for the other). ***C***, Probabilistic selection task transfer phase follows immediately after sequential task. All pairs of second-stage options are presented, and subjects are instructed to choose the stimulus with the highest chance of reward without the aid of feedback. Novel pairs of the highest reward probability option (choose 70%) and the lowest (avoid 30%) assess learning from positive and negative outcomes, respectively.

D1 versus D2 function (Cockburn et al., 2014; Collins and Frank, 2014). Immediately following the sequential task, subjects completed a transfer phase, in which their learning about these stimuli was probed (Fig. 1*C*). Specifically, they made choices between all possible pairings of the second-stage stimuli, differentially assessing choice of more rewarding versus avoidance of least rewarding options. In this phase, subjects received no feedback and were told to select the stimulus that had the highest chance of providing a reward (5 s response window). Each of the 6 pairings was presented 6 times, randomly interleaved.

Regression analyses. We fit separate multilevel logistic regressions to all choice data in the two-step sequential learning task and to the transfer phase using the lme4 package (http://cran.r-project.org/web/packages/lme4/index.html) for the R statistical language (http://www.r-project.org/). Contrasts between estimated coefficients were conducted using the esticon function in the doBy package (http://cran.r-project.org/web/packages/doBy/index.html). All within-subject predictor variables were taken as random effects (i.e., varying from subject to subject around a group mean) (Barr et al., 2013).

Two-step sequential learning task. For the 300 trials of the sequential task, we analyzed subjects' propensity to stay with or switch from the start stage choice made on the previous trial (stay coded 1, switch coded 0) as a function of the reward on the previous trial (reward coded 1, no reward coded −1), the type of state transition on the previous trial (common coded 1, uncommon coded −1), and the interaction of previous reward and previous transition type. The previous reward coefficient estimates model-free choice, whereas the interaction estimates model-based choice. In addition to these three within-subject variables, we entered as group-level predictors, *z*-scored linear genotype variables for the number of *COMT* Met alleles and *DARPP-32* T alleles, as well as the interaction of each with each of the within-subject terms in the model. Modeling the effects of both SNPs simultaneously controls for correlation in alleles across subjects. Finally, to control for population stratification in the sample, we included a racial group indicator variable (Caucasian coded 0, non-Caucasian coded 1) and its interaction with all other terms in the model. By this coding scheme, terms interacted with this variable reflect the difference of the non-Caucasian and Caucasian subsets, and the remaining terms reflect estimates for the Caucasian subset of the sample.

Transfer phase. In the transfer phase, we analyzed subjects' (putatively model-free) ability to select the stimulus with the greatest reward probability in each of the four novel pairings of the four second-stage stimuli from the sequential task (correct coded 1, incorrect coded 0). Novel pairings were grouped into those where the correct response was to choose the 70% stimulus (choose 70 trials: 70% vs 60%, 70% vs 40%), and those where the correct response was to avoid the 30% stimulus (avoid 30 trials: 30% vs 60%, 30% vs 40%), to form a trial type predictor variable (choose 70 coded 1, avoid 30 coded −1). This estimate reflects the learned ability to choose frequently rewarding actions relative to avoiding frequently nonrewarding ones. Non-zero estimates of this term are proposed to reflect differences in corticostriatal plasticity in the direct and indirect pathways (because of differential efficacy of D1- and D2-expressing striatal neurons, respectively) (Frank, 2005; Collins and Frank, 2014). As in the analysis of the two-step task, this model additionally assessed linear effects of both *COMT* and *DARPP-32*, each alone and interacted with trial type, as well as with a racial group indicator variable and its interaction with the genotype terms.

Cross task analysis. To compare relative associations of *COMT* and *DARPP-32* with model-based RL as measured in the sequential task and putative model-free RL as measured in the transfer phase, we refit the logistic regressions described above without terms for genotype or racial group. We then extracted the coefficients for each effect of interest from the separate models for each subject (model-based: previous reward × previous transition interaction from the sequential task regression; model-free: trial type term from the transfer phase regression). These coefficients were *z*-scored, and their difference entered into a linear regression with *COMT* and *DARPP-32* as independent variables (along with a racial group indicator variable and its interaction with the genotype terms).

Reinforcement learning model. The logistic regression model discussed above quantifies model-based and model-free learning in terms of their differential predictions about the effects of trial outcomes on choice in the very next trial. However, in general, RL models predict that choices are determined by values learned incrementally over multiple trials. To verify that our results were robust to the inclusion of these longer-term dependencies, in addition to the logistic regression model, we additionally fit each subject's trial-by-trial choices in this task using an RL model. This model, like the regression, assesses the contributions of model-based and model-free learning to choice, but here choices depend on values learned from the full sequence of previous rewards rather than just the last trial. We used a variant of the model of Daw et al. (2011), which has been used for many similar sequential tasks (Otto et al., 2013a, b; Doll et al., 2015a, 2015b). This model hybridizes model-free and model-based strategies, which we describe in turn.

Model-free component. The model-free approach learns a value, Q^{MF}, for each action *a* in each of the three states *s* (each state at one of the two task stages *i*). When chosen, the value of an action is updated, combining the predicted value with the outcome received on each trial *t* as follows:

$$Q^{\mathrm{MF}}_{t+1}(s_{i,t}, a_{i,t}) = (1 - \alpha_i) * Q^{\mathrm{MF}}_{t}(s_{i,t}, a_{i,t}) + r_{i,t} + Q^{\mathrm{MF}}(s_{i+1,t}, a_{i+1,t}),$$

where α_i $(0 \geq \alpha_i \geq 1)$ is the learning rate parameter estimated separately at the first and second task stages, and *r* (1, −1) is the reward or nonreward. This parameterization (Camerer and Ho, 1999) does not change the data likelihood (relative to the parameterization used by Daw et al., 2011) but facilitates group-level modeling by reducing the correlation of

α with β parameters. This update equation specializes differently at the first and second stages. Rewards are not delivered following first-stage choice, so $r_{1,t} = 0$. At the second-stage, $Q^{\mathrm{MF}}(s_{i+1,t}, a_{i+1,t}) = 0$ because there are no further states to visit in the trial.

The values of chosen first-stage actions are learned on each trial as above, and also from rewards following second-stage choices via an eligibility trace parameter, λ ($0 \geq \lambda \geq 1$) as follows:

$$Q^{\mathrm{MF}}_{t+1}(s_{1,t}, a_{1,t}) = Q^{\mathrm{MF}}_{t}(s_{1,t}, a_{1,t}) + \lambda * (r_{i,t} - Q^{\mathrm{MF}}(s_{2,t}, a_{2,t})).$$

The eligibility trace is set to 0 between episodes, such that its effects do not carry from trial to trial.

As in previous reports (Otto et al., 2013b), we decayed the Q values of unchosen actions through multiplication by $1 - \alpha$. Apart from providing a better fit to choice data across many studies (Lau and Glimcher, 2005; Ito and Doya, 2009), this ensures that, as α approaches 1, the model more closely corresponds to the regression model described above.

Model-based component. The model-based component learns a transition function (which maps state-action pairs to a probability distribution over the subsequent state) together with second-stage reward values. The values of the first-stage actions are prospectively computed, weighting the second-stage values to which they lead by their learned transition probabilities.

Following a previous report (Otto et al., 2013a), transition learning was modeled via Bayesian estimation. Actions (a_A, a_B) in the first-stage state (s_A) stochastically cause transitions to second-stage states (s_B, s_C). The probabilities of these transitions are learned over experience, beginning with a uniform Beta prior over transition probabilities. The probability of transition to second-stage s_B following action a_A in the first stage (s_A) on trial t is as follows:

$$P(s_B|s_A, a_A) = (1 + N_{AB})/(2 + N_{AC} + N_{AB})$$

where N_{AB} is the number of transitions so far experienced from state s_A to s_B following action a_A, and N_{AC} is the count of those from s_A to s_C following a_A. Transition probabilities following first-stage action a_B, and those to second-stage state s_C following actions a_A and a_B are updated similarly.

The model-based approach uses this transition function to compute first-stage values, Q^{MB}, for each action, a_j, on each trial as follows:

$$Q^{\mathrm{MB}}(s_A, a_j) = P(s_B|s_A, a_j)\max_{a \epsilon \{a_A, a_B\}} Q^{\mathrm{MF}}(s_B, a) + P(s_C|s_A, a_j)\max_{a \epsilon \{a_A, a_B\}} Q^{\mathrm{MF}}(s_C, a).$$

These model-based values arise from second-stage model-free estimates. With no further task stages in a trial, these approaches are identical at the second stage.

Choice rule. We use a softmax choice rule to connect values to choices. This rule assigns a choice probability to each action, reflecting a combination of the model-based and model-free components weighted by separate inverse temperature parameters. At the first stage, the probability of choosing action a on trial t is computed as follows:

$$P(a_{1,t} = a|s_{1,t}) = \frac{\exp\,(\beta^{\mathrm{MB}}Q^{\mathrm{MB}}(s_{1,t}, a) + \beta^{\mathrm{MF}}Q^{\mathrm{MF}}(s_{1,t}, a) + p \cdot rep(a))}{\sum_{a'} \exp\,(\beta^{\mathrm{MB}}Q^{\mathrm{MB}}(s_{1,t}, a') + \beta^{\mathrm{MF}}Q^{\mathrm{MF}}(s_{1,t}, a') + p \cdot rep(a'))}$$

The free inverse temperature parameters β^{MB} and β^{MF} control the degree to which choices follow from the model-based and model-free action values, respectively. This rule also features a perseveration parameter p, which captures the tendency to repeat actions regardless of choice. The application of this parameter is controlled by indicator function *rep(a)*, which is set to 1 if a is a first-stage action, which matches the first-stage action chosen in the previous trial (*rep(a)* = 0 otherwise). At the second stage, the model-based and model-free values are identical and action probabilities are computed as follows:

$$P(a_{2,t} = a|s_{2,t}) = \frac{\exp\,(\beta^{S2}Q^{\mathrm{MF}}(s_{2,t}, a))}{\sum_{a'} \exp\,(\beta^{S2}Q^{\mathrm{MF}}(s_{2,t}\ a'))}$$

Reinforcement learning model estimation and parameter inference. For each subject, we estimated the individual parameters of the RL model by maximizing the log posterior likelihood of the choice data conditioned on the rewards obtained (using multiple random starting points for parameters to escape local maxima). We used weakly informative priors from previous reports (Daw et al., 2011; Doll et al., 2015a) consistent with commonly observed estimates and ensuring smooth parameter boundaries. These priors were $\gamma(1.2, \text{scale} = 5)$ for softmax temperatures ($\beta^{\mathrm{MB}}, \beta^{\mathrm{MF}}, \beta^{S2}$), $\beta(1.1, 1.1)$ for parameters ranging between 0 and 1 (α_1, α_2, λ), and Normal(0, 1) for perseveration parameter p. To assess the relationship of genetic variability to model-based and model-free RL, we entered β^{MB} and β^{MF} estimates into separate linear regressions, each with *COMT* and *DARPP-32* genotypes as independent variables (together with a covariate for racial group interacted with all terms). This two-stage "summary statistics" strategy of testing group-level variation in individual-level estimates is a robust method for testing population-level effects in a mixed-effects model, parallel to the generalized linear mixed model we estimate for the logistic regression (Holmes and Friston, 1998; Friston et al., 2005).

Model comparison. To assess whether the hybrid RL model described above best described the data, we compared it with the simpler component models. To compare the three models (model-based, model-free, and the hybrid model), we computed the Laplace approximation of the model evidence (MacKay, 2003), E_m as follows:

$$E_m \approx \log p(\hat{\theta}_m) + \log p(c_{1:T} \mid \hat{\theta}_m) + \frac{1}{2} G_m \log 2\pi - \frac{1}{2} \log |H_m|$$

Where $p(\hat{\theta}_m)$ is the value of the prior at the maximum a posteriori parameter values, $p(c_{1:T} \mid \hat{\theta}_m)$ is the likelihood of the choices across trials 1 to T, G_m is the number of model parameters, and H_m is the determinant of the Hessian matrix evaluated at the posterior parameter values. This evidence, computed for each subject, was used to compare model fits at the group level by comparing summed evidences, and by the Bayesian model selection procedure of Stephan et al. (2009), which treats model identity as a random effect. This latter procedure produces a posterior probability (exceedance probability, xp_m) that each model is the most likely generative model across subjects among all those compared.

Results

Two-step sequential learning task

We investigated whether genetically indexed prefrontal and striatal DA function were associated with model-based and model-free choice strategy in a sequential learning task that dissociates these decision-making approaches (Daw et al., 2011). We took the Val^{158}Met polymorphism within the *COMT* gene as an index of prefrontal DA. *COMT* is an enzyme that catabolizes DA (Männistö and Kaakkola, 1999) and has been shown to impact prefrontal DA levels (Gogos et al., 1998; Tunbridge et al., 2004; Lapish et al., 2009), with little or no effect on striatal DA (Maj et al., 1990; Acquas et al., 1992; Li et al., 1998). The Val^{158}Met polymorphism is associated with variation in *COMT* activity (Lachman et al., 1996), with Met allele carriers showing less efficient *COMT* activity, and hence higher DA levels as reflected by differences in cortical DA binding (Slifstein et al., 2008). This genetic index of prefrontal DA levels has been implicated in behavioral and neural measures of working memory (Egan et al., 2001; Tunbridge et al., 2004; de Frias et al., 2010), consistent with DA effects on the sustained cellular activation thought to underlie working memory (Durstewitz and Seamans, 2008). We used a polymorphism within the *PPP1R1B* gene coding for *DARPP-32* as an index of striatal DA function. *DARPP-32* is a protein that is very densely expressed in the striatum and only weakly expressed in prefrontal cortex (Berger et al., 1990; Schalling et al., 1990; Ouimet et al., 1992; Meyer-Lindenberg et al., 2007). This intercellular protein accumulates following D1 receptor stimulation

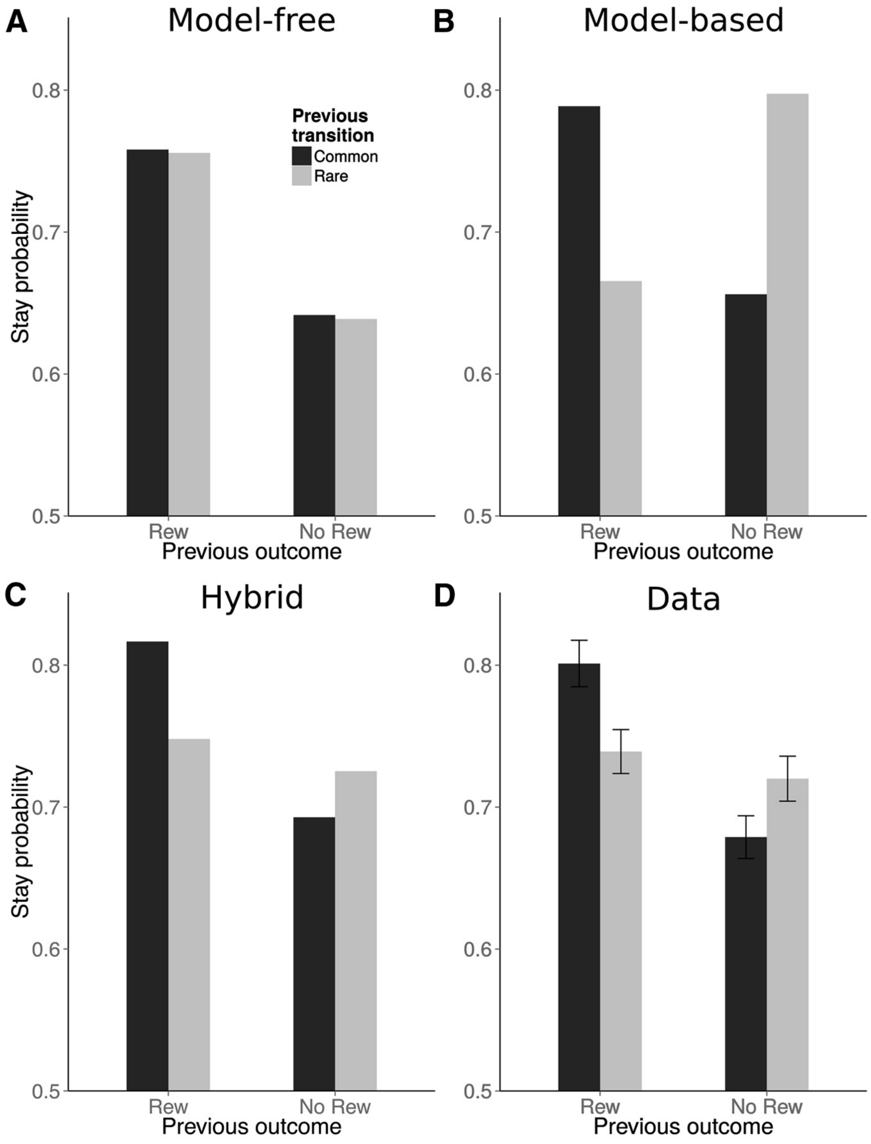

Figure 2. Model predictions and human subject data. Tendency to stay with (or switch from) the first-stage choice made on the previous trial plotted as a function of the experienced reward (Rew = reward, No Rew = no reward) and transition type on the previous trial. ***A***, The model-free strategy predicts increased stay behavior following rewards, regardless of transition type. ***B***, The model-based strategy prospectively considers reward and transition probability, thus predicting that rare transitions should affect the value first-stage action that was not chosen in the previous trial (producing an interaction between reward and transition type). ***C***, The hybrid model captures hallmarks of both strategies. Model predictions derived from simulations using the mean of the best fit parameters from human subject data. However, the predictions hold over arbitrary parameter settings. ***D***, Human data (Caucasian subset) mirror the hybrid model, showing evidence of both model-based and model-free strategies. Error bars indicate SEM.

with drugs or physiological reward and is necessary for synaptic plasticity (Calabresi et al., 2000; Valjent et al., 2005; Stipanovich et al., 2008). In putatively model-free RL tasks, the SNP we assess has been repeatedly associated with learning to select rewarding stimuli relative to avoiding nonrewarding ones (Frank et al., 2007; Doll et al., 2011; Cockburn et al., 2014). This finding is predicted by the model-free learning mechanism in a neural network model of striatum (Frank et al., 2004) and is consistent with the opposing roles of *DARPP-32* in DA-mediated synaptic plasticity in D1- and D2-expressing striatal cells (Svenningsson et al., 2004; Bateup et al., 2008).

In the two-stage sequential learning task, subjects faced a sequence of choices between two options (Fig. 1*A*). Choice between options in the first-stage state led to a second set of options in one of two second-stage states. This transition between states was stochastic, such that one of the first-stage options frequently (on 70% of choices) led to one of the second-stage states, and infrequently to the other (on 30% of choices). The other first-stage option had the reverse relationship with the second-stage states. Choice between options in the second-stage states produced reward with a slowly and randomly diffusing probability (Fig. 1*B*). These noisy transitions, together with drifting reward probabilities, required subjects to repeatedly adjust their responses to maximize earned rewards.

Model-based and model-free approaches make different behavioral predictions following outcomes on rare (30%) relative to common (70%) transitions in this task (Fig. 2*A*,*B*). A model-free approach increments the value of choices based on the outcomes that follow, regardless of the transitions experienced. Thus, if a first-stage choice is followed by a rare transition, and, ultimately, a reward, the value of the first-stage choice that started this sequence will be increased. As a result, the model-free prediction is to stay with the previous first-stage choice on the next trial (this choice is unlikely to transition to the second stage visited in the previous trial). In contrast, a model-based approach stores a representation of the noisy task structure, and uses it to prospectively plan the best choices to make. Under this strategy, a first-stage choice is made by considering the estimated rewards in the second-stage states weighted by the probability of transitioning to those states through a first-stage choice. Following a reward on a rare transition, the model-based prediction is to switch from the previous first-stage choice, and instead choose the option more likely to transition to the second-stage state that produced reward on the last trial.

To investigate genetic associations with model-based and model-free learning, we fit a multilevel logistic regression to subject data. This regression assesses the tendency to stay with or switch from the previous first-stage choice as a function of events in the previous trial, capturing the core distinguishing features of the model-based and model-free strategies (Fig. 2). Of interest are the effects of previous reward (assessing model-free choice), whether those reward effects are moderated by previous transition type (assessing model-based choice), and whether either of these effects is moderated by genotype. On average (genotypes *z*-scored), choice was influenced by the previous reward (estimate = 0.24, $z = 9.42$, $p < 2 \times 10^{-16}$), and this effect differed by the transition type on the previous trial (previous reward × previous transition interaction, estimate = 0.16, $z = 6.57$, $p = 5 \times 10^{-11}$), indicating evidence of both model-free and model-based choice, respectively (Fig. 2*D*).

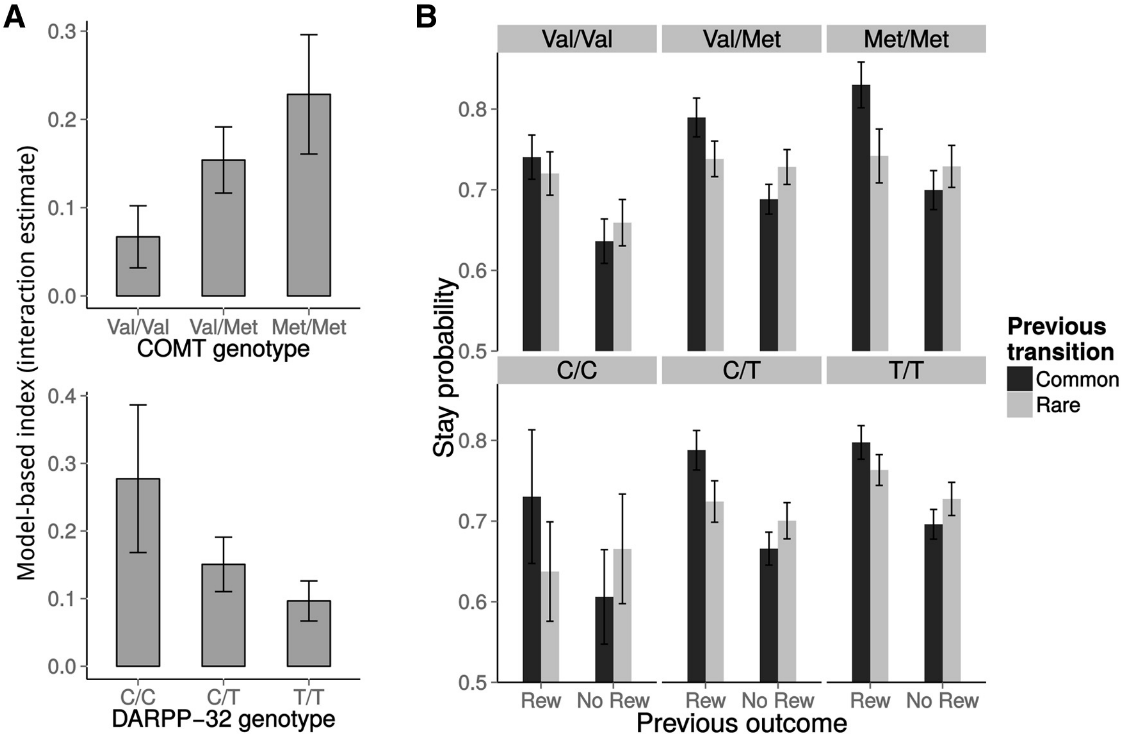

Figure 3. ***A***, Logistic regression coefficients reflecting the degree to which subjects' choices were model-based. Bars represent the interaction of previous reward (Rew = reward, No Rew = no reward) and previous transition type estimated for each genotype in Caucasian subset. Top, Model-based choice increases with *COMT* Met alleles (linear effect: $p = 0.01$). Bottom, Negative relationship of model-based choice with *DARPP-32* T alleles was not significant (linear effect: $p = 0.1$). ***B***, Sequential task mean choice proportions by genotype for Caucasian subset (top row: *COMT*; bottom row: *DARPP-32*). Error bars indicate SEM.

Table 1. Model comparisons for model-based, model-free, and hybrid models (the latter blending the first two)[a]

Model	Parameters	Evidence	*xp*
Model-based	4	61,689	0
Model-free	6	59,180	0
Hybrid	7	58,334	1

[a] For each model, table displays the number of parameters, the Laplace approximation of the model evidence (smaller numbers indicate better fit), and the Bayesian exceedance probability (*xp*) that each model is the most common across subjects among the three models.

We predicted that *COMT* Met alleles, which are associated with increased prefrontal DA and putatively enhanced computational capacity, would also reflect the degree to which behavior was model-based. Consistent with this prediction, the size of the model-based interaction term was positively correlated with the number of *COMT* Met alleles (Fig. 3; *COMT* × previous reward × previous transition interaction estimate = 0.077, $z = 2.4$, $p = 0.0154$). *COMT* did not relate significantly to the model-free term (*COMT* × previous reward estimate = 0.001, $z = 0.047$, $p = 0.96$), and the difference in *COMT* effect sizes on model-based versus model-free choice only trended toward significance (estimate = 0.075, $\chi^2 = 2.92$, $p = 0.09$).

We next asked whether *DARPP-32* genotype, which is associated with DA-mediated synaptic plasticity in striatum through opposite effects in D1 versus D2 populations (Calabresi et al., 2000; Lindskog et al., 2006; Bateup et al., 2008; Stipanovich et al., 2008; Doll et al., 2011; Cockburn et al., 2014), covaried with either strategy. We reasoned that, if D1-mediated striatal plasticity drives learning in the sequential task, *DARPP-32* T alleles would relate to model-free behavior. We observed no such relationship (*DARPP-32* × previous reward estimate = 0.01, $z = 0.37$, $p = 0.71$). Nor did *DARPP-32* relate significantly to model-based choice (Fig. 3; *DARPP-32* × previous reward × previous transition interaction: estimate = −0.062, $t = -1.66$, $p = 0.097$). Further, *DARPP-32* did not differentially associate with model-based or model-free choice (estimate = −0.08, $\chi^2 = 2.17$, $p = 0.14$).

A *post hoc* comparison showed that the effect of *COMT* Met alleles on model-based choice was larger than that of *DARPP-32* T alleles (estimate = −0.14, $\chi^2 = 6.86$, $p = 0.009$), consistent with our hypothesis that prefrontal DA plays a dissociable role in model-based RL. The effect size of these SNPs on model-free choice did not differ (estimate = −0.013, $\chi^2 = 0.05$, $p = 0.82$).

To further investigate genetic associations with model-based and model-free learning, we fit an RL model that nests these approaches to each subject's sequential learning task data (model comparisons showed this hybrid model to better fit the data than either of the nested models alone; Table 1). This model considers not just the signature impact of previous trial events as in the regression analyses, but instead assesses the time-decaying influence of all previous trial events on choice in both task stages. This allows us to verify that the results reported above are not affected by the omission of these longer-term dependencies, and may show greater power to detect effects specifically related to model-based versus model-free processing once other parameters are controlled. The model, like the regression, captures the contributions of the two strategies to choice through two separate parameters: β^{MB} (model-based choice weight) and β^{MF} (model-free choice weight). We estimated these parameters, together with parameters for learning rates and perseveration (Table 2), for each subject, and examined whether β^{MB} and β^{MF} related to *COMT* or *DARPP-32* genotype with linear regressions.

We first tested the relationship of model-based choice parameter β^{MB} to genotype. In agreement with the results described

Table 2. Quantiles of RL hybrid model parameter estimates across all subjects

Parameter	25%	50%	75%
β^{MB}	0.12	0.4	1.01
β^{MF}	0.13	0.28	0.47
β^{S2}	0.024	0.073	0.15
α_1	0.065	0.19	0.37
α_2	0.05	0.22	0.56
λ	0.54	0.84	0.97
p	0.22	0.68	1.12

above, *COMT* Met alleles showed a significant positive linear relationship with model-based choice parameter β^{MB} (estimate = 0.18, $t_{(153)} = 2.15$, $p = 0.033$; Fig. 4*A*). We found that *DARPP-32* T alleles showed the opposite relationship, relating negatively to model-based choice parameter β^{MB} (estimate = −0.23, $t_{(153)} = -2.25$, $p = 0.026$; Fig. 4*B*). A *post hoc* contrast showed the *COMT* Met allele effect to be larger than the *DARPP-32* T allele effect (estimate = 0.41, $t_{(153)} = 2.88$, $p = 0.005$). One possible explanation for the negative *DARPP-32* result is that increases in the weight of model-free relative to model-based choice might manifest negatively on the model-based term rather than positively on the model-free term (because of either their relative expression in behavior or reduced reliance on the model-based strategy as a result of model-free learning). This possibility is consistent with transfer phase analysis described below, which more specifically assesses putatively model-free statistical learning of positive and negative outcomes.

In a second linear regression, we assessed genetic associations with model-free choice parameter β^{MF}. As in the results of the logistic regression analysis, we found that neither SNP related significantly to this parameter (*COMT* estimate = 0.04, $t_{(153)} = 1.66$, $p = 0.099$; *DARPP-32* estimate = 0.014, $t_{(153)} = 0.47$, $p = 0.64$; *post hoc* contrast of difference = 0.029, $t_{(153)} = 0.66$, $p = 0.5$).

These results are consistent with our hypothesis that PFC DA, which is supposedly increased in *COMT* Met allele carriers (Lachman et al., 1996; Slifstein et al., 2008), supports model-based behavior. In contrast, we did not find clear evidence that DA-mediated striatal plasticity, to the extent that it is influenced by *DARPP-32* T alleles, is associated with model-free behavior as measured by the sequential task. Instead, we observed relatively weaker evidence that *DARPP-32* T alleles relate negatively to model-based behavior, a relationship that may arise indirectly from a positive relationship with model-free behavior. To further assess whether our measures of DA function in PFC and striatum differentially associate with behavior, we turn to the transfer phase, which provides more sensitive (though putative) measures of statistical model-free learning.

Transfer phase

Unbeknownst to subjects, the reward probabilities of the second-stage states drifted to set locations over the first 150 trials and remained fixed for the subsequent 150 trials (see Materials and Methods). This design feature permitted the learned values of the four second-stage options to be probed in a subsequent transfer phase. In this phase, which immediately followed the sequential task, subjects were presented with randomly interleaved novel and familiar pairings of the second-stage stimuli and were asked to choose the stimulus with the greatest chance of reward on each trial without the aid of feedback (Fig. 1*C*).

This transfer phase probes subjects' ability to make subtle discriminations on the basis of previous learning from positive and negative outcomes. In particular, the transfer phase contains novel trials in which the most frequently rewarding stimulus (70% rewarding stimulus) should be chosen over the more neutral options, and those in which the most frequently nonrewarding stimulus should be avoided (30% rewarding stimulus) compared with the more neutral options. These separate learning measures are hypothesized to reflect, respectively, plasticity in the direct and indirect pathway via differential DA effects on D1- and D2-expressing striatal neurons (Frank, 2005; Collins and Frank, 2014), as supported by PET data (Cox et al., 2015) and genetic evidence (Frank et al., 2007; Doll et al., 2011; Cockburn et al., 2014). Although it does not explicitly dissociate model-free from model-based choice, numerous reported effects of DA measures on performance (Frank et al., 2004, 2007; Frank and O'Reilly, 2006; Jocham et al., 2011) suggest that this phase assesses the effects of prediction-error driven incremental computations of model-free learning and choice.

We assessed the accuracy of subjects' transfer phase choices in a multilevel logistic regression with trial type (choose 70% rewarding stimulus vs avoid 30% rewarding stimulus), genotype, and their interaction as independent variables. The intercept was significantly positive (estimate = 1.08, $z = 9.29$, $p = 2 \times 10^{-16}$), indicating that subjects chose accurately on average. This accuracy varied across trial types by *DARPP-32* genotype, as indicated by a significant interaction of *DARPP-32* and trial type (Fig. 5*A*; estimate = 0.44, $z = 3.26$, $p = 0.0011$). This replicates previous work (Frank et al., 2007; Doll et al., 2011) demonstrating an advantage for T allele carriers in learning putatively model-free values from positive relative to negative outcomes. No such effect was observed for *COMT* genotype (Fig. 5*B*; estimate = −0.047, $z = -0.4$, $p = 0.68$; for summary of gene-behavior effects, see Table 3), nor did either genotype relate to overall transfer phase accuracy (*COMT* estimate = 0.15, $z = 1.39$, $p = 0.16$; *DARPP-32* estimate = −0.11, $z = -0.9$, $p = 0.36$; difference = −0.26, $\chi^2 = 2.11$, $p = 0.14$). A *post hoc* comparison confirmed that the effect of *DARPP-32* T alleles on differential transfer phase accuracy across trial types was greater than that of *COMT* Met alleles (estimate = 0.49, $\chi^2 = 6.29$, $p = 0.012$).

Cross-task analysis

COMT Met alleles positively related to model-based choice in the sequential task and showed no effect in the putative model-free measure in the transfer phase. *DARPP-32* T alleles showed relatively weaker evidence for a negative relationship with model-based choice, and a positive effect on model-free transfer phase behavior. We assessed these apparent differences in genetic associations with model-based and model-free RL across the two tasks by directly comparing them. Specifically, we entered the difference of the *z*-scored estimates from the two logistic regression analyses into a linear regression with genotypes as independent variables (see Materials and Methods). *COMT* Met alleles showed a marginal positive relationship with this difference (estimate = 0.25, $t_{(152)} = 1.86$, $p = 0.064$), whereas *DARPP-32* T alleles were negatively associated (estimate = −0.54, $t_{(152)} = -2.28$, $p = 0.0009$). These results suggest that the *DARPP-32* relationship to choice is larger for the model-free than the model-based strategy, and that the reverse association holds for *COMT* (albeit marginally). Finally, a *post hoc* contrast showed that these associations of genes and choice differed significantly from one another in direction (estimate = 0.8, $t_{(152)} = 3.47$, $p = 0.0007$).

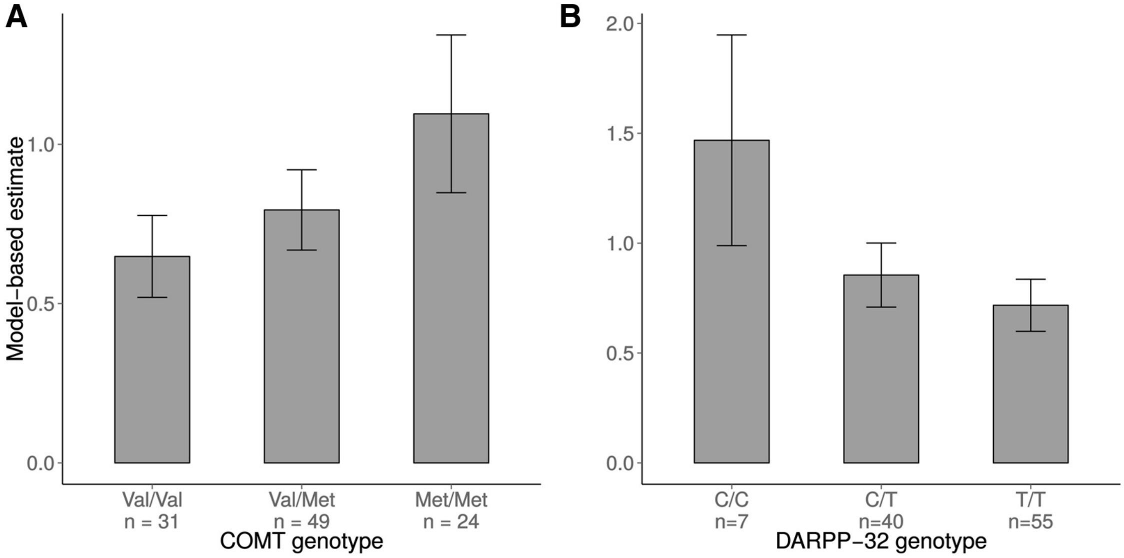

Figure 4. Model-based choice weight parameter β^{MB} estimates from computational model plotted by genotype (Caucasian subset). ***A***, Parameter β^{MB} increases with *COMT* Met alleles (linear effect: $p = 0.03$), which are putatively associated with increased DA levels in PFC. ***B***, Parameter β^{MB} decreases with *DARPP-32* T alleles (linear effect: $p = 0.03$), which are putatively associated with enhanced striatal DA-mediated learning from positive relative to negative outcomes. Error bars indicate SEM.

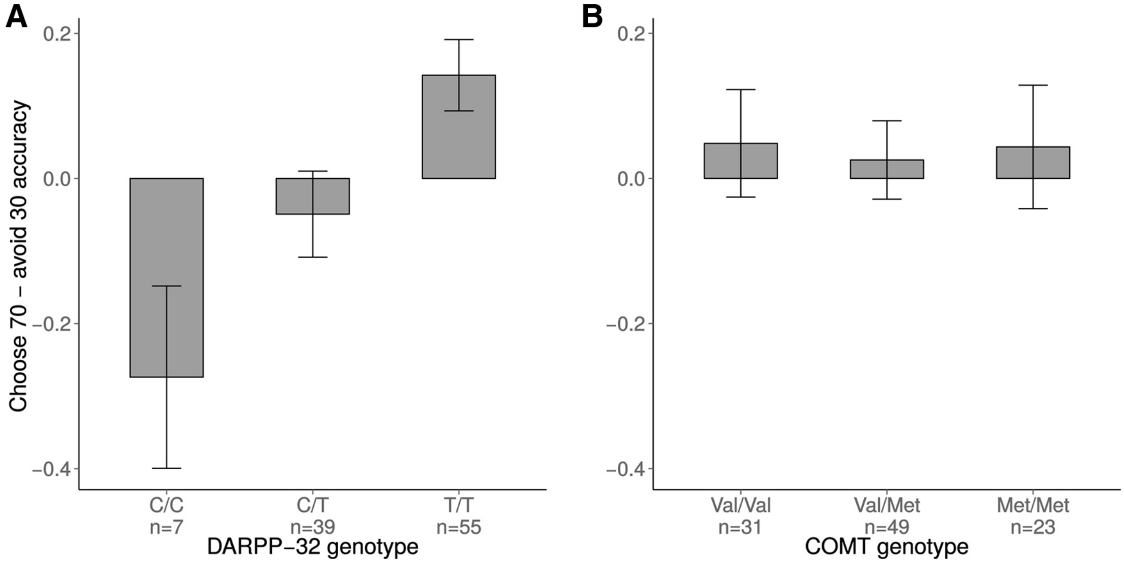

Figure 5. Probabilistic selection transfer phase accuracy by genotype for Caucasian subset. Differential accuracy in choosing the most highly rewarding (Choose 70%) versus avoiding the least rewarding (Avoid 30%) stimulus is hypothesized to reflect differential DA-mediated plasticity in the direct and indirect pathways (via opposing effects on D1- and D2-expressing striatal cells, respectively), learned by model-free RL during the sequential task. ***A***, *DARPP-32* T alleles are associated with learning from positive relative to negative outcomes (linear effect: $p = 0.001$). Error bars indicate SEM. ***B***, *COMT* genotype shows no relationship with accuracy in learning from positive and negative outcomes (linear effect: $p = 0.7$).

Sample stratification

Isolation between populations imposes nonrandom mating, producing differences in allele frequencies and linkage disequilibrium between races, thereby complicating genetic association studies with heterogeneous samples (Cardon and Palmer, 2003). In some cases, candidate SNPs have been found to exist on different racial population-specific haplotypes (Petryshen et al., 2010), highlighting the importance of controlling for population stratification in genetic association studies. Indeed, allele frequencies of the *COMT* SNP studied here differ by population (McLeod et al., 1998), and cognitive associations with *COMT* have even been found to differ across racial groups (Humphreys et al., 2014). These differences could arise for a number of reasons that follow from allele frequency and linkage disequilibrium differences across populations. For example, the magnitude and direction of DA modulation effects on cognition often depend on baseline DA levels, which in turn may be affected by other genetic factors that could covary with race.

We controlled potential stratification confounds as in previous reports (Frank et al., 2009; Doll et al., 2011; Collins and Frank, 2012; Cockburn et al., 2014), by confirming that the results held in the largest racial group in the sample. Specifically, we included a covariate indicating subjects (66 of 171) who self-identified as non-Caucasian in the regressions of behavior and RL model parameters (Caucasian coded 0, non-Caucasian coded 1). Thus, the foregoing results describe the genetic associations in the Caucasian subset. Below, we report that the regression coefficients of these covariates interacted with behavioral and genetic

Table 3. Summary of model coefficients reflecting gene-behavior associations[a]

Gene	Analysis	Effect	Effect size	Statistic	p
COMT	Logistic-1	MB	0.08	$Z = 2.4$	0.01
	Logistic-1	MF	0.001	$Z = 0.05$	0.96
	RL-1	MB	0.18	$t_{(153)} = 2.1$	0.03
	RL-2	MF	0.04	$t_{(153)} = 1.7$	0.1
	Logistic-2	MF*	−0.05	$Z = -0.4$	0.7
	Cross-task	MB-MF*	0.25	$t_{(152)} = 1.9$	0.06
DARPP-32	Logistic-1	MB	−0.06	$t_{(153)} = -1.7$	0.1
	Logistic-1	MF	0.01	$Z = 0.4$	0.7
	RL-1	MB	−0.23	$t_{(153)} = -2.2$	0.03
	RL-2	MF	0.01	$t_{(153)} = 0.5$	0.6
	Logistic-2	MF*	0.44	$Z = 3.2$	0.001
	Cross-task	MB-MF*	−0.54	$t_{(152)} = -2.3$	0.001

[a] Shown are effects summaries from five regression models. Gene, *COMT* effects oriented with Met alleles positively and *DARPP-32* with T alleles positively; Analysis, Logistic-1 = regression model of stay/switch behavior in sequential task; RL-1 = linear regression model of RL parameter β^{MB} estimated from computational model fits to subject behavior; RL-2 = regression model of parameter β^{MF}; Logistic-2 = model of transfer phase accuracy; Cross-task = regression of stay/switch estimates reflecting model-based RL relative to estimates reflecting transfer phase accuracy; Effect, MB = model-based; MF = model-free; MF* = putative model-free in transfer phase; Effect size, regression coefficients.

variables. These coefficients capture the difference in effects between the Caucasian and non-Caucasian subsets of the sample.

In the logistic regression analysis of behavior, we observed no differences in model-based or model-free choice between racial groups on average (previous reward difference = −0.047, $z = -1.14$, $p = 0.25$; previous reward × previous transition difference = −0.018, $z = -0.045$, $p = 0.65$). Nor did we observe any significant differences in genetic associations with these variables in the non-Caucasian subset (*COMT* × previous reward difference = 0.03, $z = 0.7$, $p = 0.47$; *COMT* × previous reward × previous transition difference = −0.06, $z = 1.6$, $p = 0.1$; *DARPP-32* × previous reward difference = −0.02, $z = -0.522$, $p = 0.6$; *DARPP-32* × previous reward × previous transition difference = 0.07, $z = 1.8$, $p = 0.07$).

In the analysis of RL model parameters, we observed no significant differences in β^{MF} or β^{MB} parameters between racial groups (β^{MF} difference = −0.01, $t_{(153)} = -0.24$, $p = 0.81$; β^{MB} difference = −0.13, $t_{(153)} = -0.81$, $p = 0.42$). Genetic associations of these parameters with *COMT* genotype did not differ significantly between racial groups (*COMT* × β^{MF} difference = −0.03, $t_{(153)} = -0.67$, $p = 0.5$; *COMT* × β^{MB} difference = −0.03, $t_{(153)} = -0.24$, $p = 0.81$). Nor did we observe significant differences in the association of these parameters with *DARPP-32* genotype (*DARPP-32* × β^{MF} difference = −0.05, $t_{(153)} = -1$, $p = 0.3$; *DARPP-32* × β^{MB} difference = 0.28, $t_{(153)} = 1.8$, $p = 0.07$).

Similarly, we observed no significant differences in the Caucasian and non-Caucasian subsets in the logistic regression model of transfer phase performance (accuracy difference = −0.24, $z = -1.2$, $p = 0.22$). Nor did we find significant differences in the relationship of genotype to accuracy across groups (*COMT* difference = −0.1, $z = -0.58$, $p = 0.56$; *COMT* × trial type difference = −0.32, $z = -1.6$, $p = 0.1$; *DARPP-32* difference = 0.1, $z = 0.5$, $p = 0.61$; *DARPP-32* × trial type difference = −0.05, $z = -0.27$, $p = 0.79$).

Discussion

In this study, we used a genetic approach to investigate whether the variations in prefrontal and striatal DA are differentially associated with model-based versus model-free RL. Recent work has shown that model-based control is associated with increased DA levels (Wunderlich et al., 2012; Deserno et al., 2015; Sharp et al., 2015). However, these studies did not permit comparison of DA function in the striatum versus prefrontal cortex, leaving open the question of whether either region is preferentially involved. Moreover, prominent studies link reward learning to dopaminergic effects on striatal plasticity, but it is unclear to what extent this mechanism specifically underlies model-free (vs model-based) RL. Here, we assessed covariation of model-based and model-free choice with two SNPs differentially involved in prefrontal and striatal DA function. *COMT* has been shown to impact prefrontal DA levels, with negligible effects in striatum (Gogos et al., 1998; Sesack et al., 1998; Huotari et al., 2002; Matsumoto et al., 2003; Tunbridge et al., 2004), whereas *DARPP-32* is far more enriched in striatum, accumulates as a function of reward, and is necessary for synaptic plasticity and reward learning (Stipanovich et al., 2008).

Consistent with the view that the prefrontal cortex supports DA-dependent working memory function (Sawaguchi et al., 1990; Sawaguchi and Goldman-Rakic, 1991; Durstewitz and Seamans, 2008), which in turn supports model-based computation (Otto et al., 2013a, b; Smittenaar et al., 2013), we found that prefrontal DA, as indexed by *COMT* Met alleles, correlated positively with model-based choice. In contrast, we observed that *DARPP-32* T alleles, as one modulator of DA-mediated striatal plasticity, related positively to signatures of striatal-dependent, ostensibly model-free learning as assessed by choices in the transfer phase. In the sequential task, *DARPP-32* T alleles related negatively to the model-based RL parameter. On the view that behavior reflects a weighted combination of these strategies, we interpret this effect as an increase in model-free relative to model-based choice (this effect was not significant in the analogous logistic regression analysis and should be interpreted cautiously).

The association of prefrontal DA with model-based learning is readily incorporated into theory and with extant empirical data (Daw et al., 2005; Wunderlich et al., 2012; Deserno et al., 2015). The working memory faculties of this region (Funahashi and Kubota, 1994) make it an auspicious candidate for computationally heavy model-based methods (Daw et al., 2005). Indeed, evidence links both lateral PFC function (Smittenaar et al., 2013) and working memory (Otto et al., 2013a, b) to model-based decision-making. Working memory, in turn, is modulated by PFC DA levels (Sawaguchi and Goldman-Rakic, 1991; Durstewitz and Seamans, 2008). The current data indicate that supposed increases in prefrontal DA across subjects as indexed by *COMT* Met alleles relate positively to model-based learning. This finding complements other observations of apparent *COMT* involvement in working memory during RL tasks (Frank et al., 2007; but see Krugel et al., 2009; Collins and Frank, 2012), and extends earlier work on DA and model-based learning (Wunderlich et al., 2012; Deserno et al., 2015). Whereas previous work showed that model-based choice increases with brain-wide DA, the current study supports the interpretation that this may have reflected influences on PFC DA. Although our results do not rule out possible involvement of prefrontal DA in model-free RL, they show affirmative evidence of involvement in model-based RL. Further, the *COMT* relationship with model-free transfer phase choice was marginally smaller than the relationship with model-based choice and significantly smaller than the *DARPP-32* association with model-free transfer choice, bolstering the view that prefrontal DA has a specific role in model-based learning.

In contrast to the *COMT* effects, we found evidence that DA-mediated striatal plasticity, as indexed by *DARPP-32*, was involved in putative model-free learning. We observed clear effects of *DARPP-32* in the transfer phase, replicating previous findings that T alleles positively associate with learning from positive relative to negative outcomes (Frank et al., 2007; Doll et al., 2011; Cockburn et al., 2014). Although the transfer phase does not explicitly separate model-based from model-free choice, theory predicts that this result stems from model-free learning implemented by asymmetrical effects of DA prediction errors on striatal D1- and D2-expressing neurons in the direct and indirect pathways (Frank, 2005; Collins and Frank, 2014). To the extent that model-free behavior in the sequential task follows from reward-driven plasticity in the direct pathway (via D1 receptor activation), we expect increases in plasticity to be accompanied by increases in model-free behavior. However, RL model fits revealed that *DARPP-32* T alleles were not positively associated with model-free parameter β^{MF} but instead negatively associated with model-based parameter β^{MB}. If these strategies are expressed relative to one another in behavior, increases in model-free choice should be accompanied by decreases in model-based choice, and vice-versa. By this view, our results suggest that *DARPP-32* T alleles covary positively with model-free behavior on both tasks. Although plausible, this interpretation should be considered in light of limitations of the sequential task, and of the *DARPP-32* index, which we discuss below.

One possible contribution to the discrepancy of *DARPP-32* results across tasks is that the sequential task better measures model-based than model-free behavior. The logic of the task is to isolate transition-based switching behavior (a signature of model-based reasoning), whereas the remaining learning might arise from some mix of explicit switching strategies with true incremental, implicit model-free learning. Moreover, the latter component might be poorly detected (compared with the transfer phase task) because it fails to separately consider striatal plasticity in the direct and indirect pathways (affecting learning from positive and negative outcomes, respectively). Also, model-free learning may accumulate slowly over time, and so might be less sensitively detected in a dynamic task of this sort (relative to the aggregate transfer phase learning measure). Indeed, studies using this sequential task, as here, overwhelmingly show group effects to load on model-based, not model-free, regression coefficients (Wunderlich et al., 2012; Otto et al., 2013a, b; Voon et al., 2014; Gillan et al., 2015). Further, the association of *DARPP-32* with our model-free measure in the transfer phase was comparatively large and clear. Future work should seek to develop sequential tasks that better capture both model-free and model-based RL by discriminating learning from positive and negative outcomes, as in the composite task used here.

Another interpretational complexity attending the *DARPP-32* result is that this SNP is thought to be a relative measure of striatal plasticity in the direct and indirect pathways. Specifically, the SNP we assess affects overall *DARPP-32* mRNA expression (Meyer-Lindenberg et al., 2007), which has opposing phosphorylation effects following D1 and D2 receptor stimulation (Svenningsson et al., 2004; Bateup et al., 2008). Increases in DA-mediated synaptic potentiation in the D1 pathway should thus be accompanied by depression in the D2 pathway. These bidirectional effects at the cellular level are mirrored in behavior by the relative effect of *DARPP-32* genotype on accuracy in choosing rewarding stimuli relative to avoiding nonrewarding ones in the transfer phase here and in prior studies. As such, the negative association of *DARPP-32* T alleles with model-based behavior in the sequential task could be interpreted as a positive association of *DARPP-32* C alleles with model-based behavior, although the computational mechanism underlying this latter relationship is unclear.

Overall, our findings that model-based learning in the sequential decision task is, if anything, negatively associated with *DARPP-32* T alleles, whereas learning from positive outcomes in the probabilistic selection task is positively associated with this measure, support the hypothesis that dopaminergic effects on model-based and model-free learning reflect partially dissociable contributions via prefrontal and striatal mechanisms, respectively. Our findings also suggest that the sequential decision and probabilistic selection tasks preferentially capture these two sorts of learning: model-based and model-free, respectively. That said, numerous studies do point to a role for striatum in model-based RL (Yin et al., 2005; Daw et al., 2011; Simon and Daw, 2011). These studies typically lack DA measurements, but one recent report (Deserno et al., 2015) showed that striatal presynaptic DA, measured by FDOPA uptake, was positively associated with model-based behavior. The authors focused on striatum because their PET technique precluded measurement of cortical dopamine. Thus, the apparent regional specificity of their reported effect might be due to the regional specificity of their measurement rather than to the site of dopaminergic action. That is, striatal FDOPA uptake might only be a proxy for more global dopaminergic variations, with PFC as the key locus. Indeed, Deserno et al. (2015) found that striatal DA levels measured with PET correlated positively with model-based coding in PFC measured with fMRI. This interpretation of the Deserno et al. (2015) study is consonant with the *COMT* findings observed here.

In conclusion, these results affirm a role for prefrontal DA in model-based RL, extending previous findings. This role is consistent with the view that prefrontal working memory function supports the intensive computations required by a model-based strategy. The current results also implicate DA-mediated striatal plasticity in model-free learning but suggest that this plasticity might not contribute to model-based learning.

References

Acquas E, Carboni E, de Ree RH, Da Prada M, Di Chiara G (1992) Extracellular concentrations of dopamine and metabolites in the rat caudate after oral administration of a novel catechol-O-methyltransferase inhibitor Ro 40–7592. J Neurochem 59:326–330. CrossRef Medline

Barr DJ, Levy R, Scheepers C, Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. J Mem Lang 68:3. CrossRef Medline

Bateup HS, Svenningsson P, Kuroiwa M, Gong S, Nishi A, Heintz N, Greengard P (2008) Cell type-specific regulation of DARPP-32 phosphorylation by psychostimulant and antipsychotic drugs. Nat Neurosci 11: 932–939. CrossRef Medline

Berger B, Febvret A, Greengard P, Goldman-Rakic PS (1990) DARPP-32, a phosphoprotein enriched in dopaminoceptive neurons bearing dopamine D1 receptors: distribution in the cerebral cortex of the newborn and adult rhesus monkey. J Comp Neurol 299:327–348. CrossRef Medline

Calabresi P, Gubellini P, Centonze D, Picconi B, Bernardi G, Chergui K, Svenningsson P, Fienberg AA, Greengard P (2000) Dopamine and cAMP-regulated phosphoprotein 32 kDa controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. J Neurosci 20:8443–8451. Medline

Camerer C, Ho TH (1999) Experience-weighted attraction learning in normal form games. Econometrica 67:827–874. CrossRef Medline

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598–604. CrossRef Medline

Cockburn J, Collins AG, Frank MJ (2014) A reinforcement learning mechanism responsible for the valuation of free choice. Neuron 83:551–557. CrossRef Medline

Collins AG, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35:1024–1035. CrossRef Medline

Collins AG, Frank MJ (2014) Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev 121:337–366. CrossRef Medline

Cox SM, Frank MJ, Larcher K, Fellows LK, Clark CA, Leyton M, Dagher A

(2015) Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes. Neuroimage 109:95–101. CrossRef Medline

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

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215. CrossRef Medline

de Frias CM, Marklund P, Eriksson E, Larsson A, Oman L, Annerbrink K, Bäckman L, Nilsson LG, Nyberg L (2010) Influence of COMT gene polymorphism on fMRI-assessed sustained and transient activity during a working memory task. J Cogn Neurosci 22:1614–1622. CrossRef Medline

Huys QJ, Boehme R, Buchert R, Heinze HJ, Grace AA, Dolan RJ, Heinz A, Schlagenhauf F (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc Natl Acad Sci U S A 112:1595–1600. CrossRef Medline

Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80: 312–325. CrossRef Medline

Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J Neurosci 31: 6188–6198. CrossRef Medline

Doll BB, Simon DA, Daw ND (2012) The ubiquity of model-based reinforcement learning. Curr Opin Neurobiol 22:1075–1081. CrossRef Medline

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015a) Model-based choices involve prospective neural activity. Nat Neurosci 18:1–9. CrossRef Medline

Doll BB, Shohamy D, Daw ND (2015b) Multiple memory systems as substrates for multiple decision systems. Neurobiol Learn Mem 117:4–13. CrossRef Medline

Durstewitz D, Seamans JK (2008) The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-O-methyltransferase genotypes and schizophrenia. Biol Psychiatry 64:739–749. CrossRef Medline

Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE, Goldman D, Weinberger DR (2001) Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. Proc Natl Acad Sci U S A 98:6917–6922. CrossRef Medline

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

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

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

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

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12:1062–1068. CrossRef Medline

Friston KJ, Stephan KE, Lund TE, Morcom A, Kiebel S (2005) Mixed-effects and fMRI studies. Neuroimage 24:244–252. CrossRef Medline

Funahashi S, Kubota K (1994) Working memory and prefrontal cortex. Neurosci Res 21:1–11. CrossRef Medline

Gillan CM, Otto AR, Phelps EA, Daw ND (2015) Model-based learning protects against forming habits. Cogn Affect Behav Neurosci 15:523–536. CrossRef Medline

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

Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A 108:1–8. CrossRef Medline

Gogos JA, Morgan M, Luine V, Santha M, Ogawa S, Pfaff D, Karayiorgou M (1998) Catechol-O-methyltransferase-deficient mice exhibit sexually dimorphic changes in catecholamine levels and behavior. Proc Natl Acad Sci U S A 95:9991–9996. CrossRef Medline

Holmes A, Friston K (1998) Generalisability, random effects and population inference. Neuroimage 7:754.

Humphreys KL, Scheeringa MS, Drury SS (2014) Race moderates the association of catechol-O-methyltransferase genotype and posttraumatic stress disorder in preschool children. J Child Adolesc Psychopharmacol 24:454–457. CrossRef Medline

Huotari M, Gogos JA, Karayiorgou M, Koponen O, Forsberg M, Raasmaja A, Hyttinen J, Männistö PT (2002) Brain catecholamine metabolism in catechol-O-methyltransferase (COMT)-deficient mice. Eur J Neurosci 15:246–256. CrossRef Medline

Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci 29:9861–9874. CrossRef Medline

Jocham G, Klein TA, Ullsperger M (2011) Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci 31:1606–1613. CrossRef Medline

Kravitz AV, Tye LD, Kreitzer AC (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci 15:816–818. CrossRef Medline

Krugel LK, Biele G, Mohr PN, Li SC, Heekeren HR (2009) Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc Natl Acad Sci U S A 106:17951–17956. CrossRef Medline

Lachman HM, Papolos DF, Saito T, Yu YM, Szumlanski CL, Weinshilboum RM (1996) Human catechol-O-methyltransferase pharmacogenetics: description of a functional polymorphism and its potential application to neuropsychiatric disorders. Pharmacogenetics 6:243–250. CrossRef Medline

Lapish CC, Ahn S, Evangelista LM, So K, Seamans JK, Phillips AG (2009) Tolcapone enhances food-evoked dopamine efflux and executive memory processes mediated by the rat prefrontal cortex. Psychopharmacology (Berl) 202:521–530. CrossRef Medline

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

Li YH, Wirth T, Huotari M, Laitinen K, MacDonald E, Männistö PT (1998) No change of brain extracellular catecholamine levels after acute catechol-O-methyltransferase inhibition: a microdialysis study in anaesthetized rats. Eur J Pharmacol 356:127–137. CrossRef Medline

Lindskog M, Kim M, Wikström MA, Blackwell KT, Kotaleski JH (2006) Transient calcium and dopamine increase PKA activity and DARPP-32 phosphorylation. PLoS Comput Biol 2:1045–1060. CrossRef Medline

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

Maj J, Rogóz Z, Skuza G, Sowińska H, Superata J (1990) Behavioural and neurochemical effects of Ro 40–7592, a new COMT inhibitor with a potential therapeutic activity in Parkinson's disease. J Neural Transm Park Dis Dement Sect 2:101–112. CrossRef Medline

Männistö PT, Kaakkola S (1999) Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors. Pharmacol Rev 51:593–628. Medline

Matsumoto M, Weickert CS, Akil M, Lipska BK, Hyde TM, Herman MM, Kleinman JE, Weinberger DR (2003) Catechol O-methyltransferase mRNA expression in human and rat brain: evidence for a role in cortical neuronal function. Neuroscience 116:127–137. CrossRef Medline

McLeod HL, Syvänen AC, Githang'a J, Indalo A, Ismail D, Dewar K, Ulmanen I, Sludden J (1998) Ethnic differences in catechol O-methyltransferase pharmacogenetics: frequency of the codon 108/158 low activity allele is lower in Kenyan than Caucasian or South-west Asian individuals. Pharmacogenetics 8:195–199. Medline

Meyer-Lindenberg A, Straub RE, Lipska BK, Verchinski BA, Goldberg T, Callicott JH, Egan MF, Huffaker SS, Mattay VS, Kolachana B, Kleinman JE, Weinberger DR (2007) Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. J Clin Invest 117:672–682. CrossRef Medline

Otto AR, Gershman SJ, Markman AB, Daw ND (2013a) The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci 24:751–761. CrossRef Medline

Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013b) Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941–20946. CrossRef Medline

Ouimet CC, LaMantia AS, Goldman-Rakic P, Rakic P, Greengard P (1992)

Immunocytochemical localization of DARPP-32, a dopamine and cyclic-AMP-regulated phosphoprotein, in the primate brain. J Comp Neurol 323:209–218. CrossRef Medline

Petryshen TL, Sabeti PC, Aldinger KA, Fry B, Fan JB, Schaffner SF, Waggoner SG, Tahl AR, Sklar P (2010) Population genetic study of the brain-derived neurotrophic factor (BDNF) gene. Mol Psychiatry 15:810–815. CrossRef Medline

Sawaguchi T, Goldman-Rakic PS (1991) D1 dopamine receptors in prefrontal cortex: involvement in working memory. Science 251:947–950. CrossRef Medline

Sawaguchi T, Matsumura M, Kubota K (1990) Catecholaminergic effects on neuronal-activity related to a delayed-response task in monkey prefrontal cortex. J Neurophysiol 63:1385–1400. Medline

Schalling M, Djurfeldt M, Hökfelt T, Ehrlich M, Kurihara T, Greengard P (1990) Distribution and cellular localization of DARPP-32 mRNA in rat brain. Brain Res Mol Brain Res 7:139–149. CrossRef Medline

Sesack SR, Hawrylak VA, Matus C, Guido MA, Levey AI (1998) Dopamine axon varicosities in the prelimbic division of the rat prefrontal cortex exhibit sparse immunoreactivity for the dopamine transporter. J Neurosci 18:2697–2708. Medline

Sharp ME, Foerde K, Daw ND, Shohamy D (2015) Dopamine selectively remediates "model-based" reward learning: a computational approach. Brain pii: awv347. CrossRef Medline

Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. J Neurosci 31:5526–5539. CrossRef Medline

Slifstein M, Kolachana B, Simpson EH, Tabares P, Cheng B, Duvall M, Frankle WG, Weinberger DR, Laruelle M, Abi-Dargham A (2008) COMT genotype predicts cortical-limbic D1 receptor availability measured with [11C]NNC112 and PET. Mol Psychiatry 13:821–827. CrossRef Medline

Smittenaar P, FitzGerald TH, Romei V, Wright ND, Dolan RJ (2013) Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron 80:914–919. CrossRef Medline

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966–973. CrossRef Medline

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004–1017. CrossRef Medline

Stipanovich A, Valjent E, Matamales M, Nishi A, Ahn JH, Maroteaux M, Bertran-Gonzalez J, Brami-Cherrier K, Enslen H, Corbillé AG, Filhol O, Nairn AC, Greengard P, Hervé D, Girault JA (2008) A phosphatase cascade by which rewarding stimuli control nucleosomal response. Nature 453:879–884. CrossRef Medline

Svenningsson P, Nishi A, Fisone G, Girault JA, Nairn AC, Greengard P (2004) DARPP-32: an integrator of neurotransmission. Annu Rev Pharmacol Toxicol 44:269–296. CrossRef Medline

Thorndike EL (1898) Animal intelligence: an experimental study of the associative processes in animals. Psychol Rev Monogr Suppl 2:1–8.

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55: 189–208. CrossRef Medline

Tunbridge EM, Bannerman DM, Sharp T, Harrison PJ (2004) Catechol-o-methyltransferase inhibition improves set-shifting performance and elevates stimulated dopamine release in the rat prefrontal cortex. J Neurosci 24:5331–5335. CrossRef Medline

Valjent E, Pascoli V, Svenningsson P, Paul S, Enslen H, Corvol JC, Stipanovich A, Caboche J, Lombroso PJ, Nairn AC, Greengard P, Hervé D, Girault JA (2005) Regulation of a protein phosphatase cascade allows convergent dopamine and glutamate signals to activate ERK in the striatum. Proc Natl Acad Sci U S A 102:491–496. CrossRef Medline

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, Schreiber LR, Gillan C, Fineberg NA, Sahakian BJ, Robbins TW, Harrison NA, Wood J, Daw ND, Dayan P, Grant JE, Bullmore ET (2014) Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry 20: 345–352. CrossRef Medline

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

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22: 513–523. CrossRef Medline