nature neuroscience

Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation

Michael J Frank[1–3], Bradley B Doll[1–3], Jen Oas-Terpstra[4] & Francisco Moreno[4]

The basal ganglia support learning to exploit decisions that have yielded positive outcomes in the past. In contrast, limited evidence implicates the prefrontal cortex in the process of making strategic exploratory decisions when the magnitude of potential outcomes is unknown. Here we examine neurogenetic contributions to individual differences in these distinct aspects of motivated human behavior, using a temporal decision-making task and computational analysis. We show that two genes controlling striatal dopamine function, *DARPP-32* (also called *PPP1R1B*) and *DRD2*, are associated with exploitative learning to adjust response times incrementally as a function of positive and negative decision outcomes. In contrast, a gene primarily controlling prefrontal dopamine function (*COMT*) is associated with a particular type of 'directed exploration', in which exploratory decisions are made in proportion to Bayesian uncertainty about whether other choices might produce outcomes that are better than the status quo. Quantitative model fits reveal that genetic factors modulate independent parameters of a reinforcement learning system.

Individuals differ in their choices and neural responses when confronted with decision uncertainty[1,2]. Some people are motivated by having achieved desirable outcomes and are driven to work harder to attain even better ones, whereas others are primarily motivated to avoid negative outcomes[3]. However, individuals often don't know which outcomes should be considered positive until they compare them to those obtained from other decision strategies (for example, do you choose to return to the same failsafe sushi restaurant or to try a new one because it might be even better?). This classic problem of whether to sample other options or maintain the current strategy for maximizing reward is known as the exploration/exploitation dilemma[4–7]. Here we examine neurogenetic contributions to exploitative and exploratory behavior.

In part, individual differences in personality variables are thought to reflect different parameters within the dopaminergic motivational system[8]. Dopaminergic genetic components that alter function in the striatum (and indirectly its interactions with frontal cortex[9]) differentiate between individuals who are more adept at learning from positive as compared to negative decision outcomes, via modulation of striatum and its interactions with frontal cortex[9–11]. Specifically, a functional polymorphism within the *DARPP-32* gene—whereby carriers of two copies of the 'T' allele (T/T carriers) show greater *DARPP-32* mRNA expression than those with at least one copy of the 'C' allele (C carriers)[9]—is predictive of 'Go learning' to reproduce behaviors that yield positive outcomes[10]. The DARPP-32 protein is highly concentrated in the striatum, is phosphorylated by D1 dopamine receptor stimulation, and is required for striatal D1 receptor–mediated synaptic plasticity and behavioral reward learning[12–14]. Although DARPP-32 is also present in D2-containing neurons, stimulation of D2 receptors dephosphorylates DARPP-32 and does not mediate its effects on reward learning[13]. Conversely, polymorphisms within the *DRD2* gene predictive of striatal D2 receptor density are associated with 'NoGo learning' to avoid behaviors that yield negative outcomes[10,11]: individuals with two copies of the *DRD2* 'T' allele (T/T carriers) have greater striatal D2 receptor density[15]. These findings converge with the notion that dopamine has a key role in reinforcement learning[16] and, in particular, that dopamine acts in the striatum to support learning from positive and negative outcomes via D1 and D2 receptors in separate neuronal striatonigral and striatopallidal populations[17,18]. The findings also converge with rodent data showing that the transition to exploitative behavior is associated with the development of highly stabilized striatal firing patterns[19].

Although the role of striatal dopamine in reinforcement exploitation is relatively well established, the neurobiological correlates of exploration are far less developed. Computational considerations suggest that an adaptive heuristic is to explore in proportion to one's uncertainty about the consequent outcomes[4,6,7,20]. Such computations might depend on neuromodulation within the prefrontal cortex (PFC)[7]. Functional neuroimaging evidence implicates anterior and orbital PFC in computations of uncertainty[2,21] and in the making of exploratory decisions in a reinforcement learning environment[6]. Further, models and experimental data suggest that orbital PFC represents reward magnitudes, which are required to compute the expected value of decisions, especially over delays[6,22–24]. At the genetic level, the gene *COMT*, encoding catechol-*O*-methyltransferase (COMT), substantially affects PFC dopamine levels and, in turn, PFC-dependent cognitive function[25]. COMT is an enzyme that breaks down dopamine; an allele of *COMT* encoding valine at chr22:18331271 (known as the 'val' allele)

[1]Departments of Cognitive & Linguistic Sciences, [2]Psychology and [3]Psychiatry, Brown Institute for Brain Science, Brown University, Providence, Rhode Island, USA. [4]Department of Psychiatry, University of Arizona, Tucson, Arizona, USA. Correspondence should be addressed to M.J.F. (michael_frank@brown.edu).

Received 13 March; accepted 28 April; published online 20 July 2009; corrected after print 9 September 2009; doi:10.1038/nn.2342

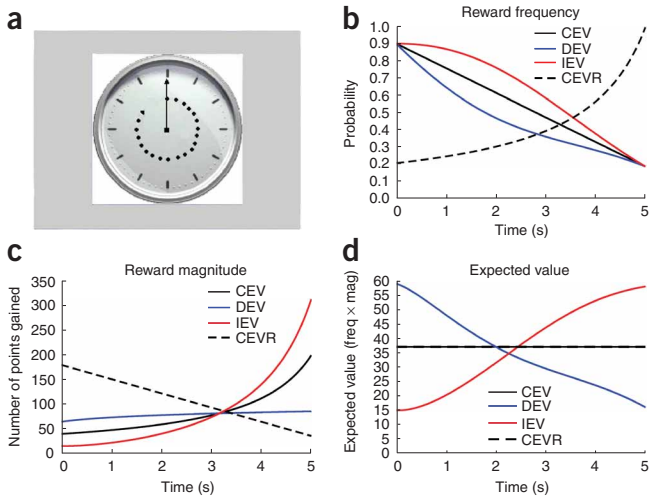

Figure 1 Task conditions: decreasing expected value (DEV), constant expected value (CEV), increasing expected value (IEV) and constant expected value–reverse (CEVR). The *x* axis corresponds to the time after onset of the clock stimulus at which the response is made. The functions are designed such that the expected value at the beginning in DEV is equal to that at the end in IEV so that at optimal performance, subjects should obtain the same average reward in both IEV and DEV. Faster responses were accompanied by longer intertrial intervals so that reward rate is roughly equalized across conditions. (**a**) Example clock-face stimulus. Each trial ended when the subject made a response or otherwise when the 5 s duration elapsed. The number of points won on the current trial was displayed. (**b**) Probability of reward occurring as a function of response time. (**c**) Reward magnitude (contingent on probability in **b**). (**d**) Expected value across trials for each time point. Note that CEV and CEVR have the same EV.

is associated with greater enzymatic efficacy, and therefore lower PFC dopamine levels, than a methionine-encoding ('met') allele. The enzyme has a comparatively minor role in striatum owing to its relatively sparse expression and to the presence of potent dopamine transporters and autoreceptors[25–28].

We assessed these motivational components, including exploitation, exploration, and probability versus magnitude learning, within a single "temporal utility integration task"[29]. We hypothesized that genetic markers of striatal dopaminergic function (*DARPP-32* and *DRD2*) would be predictive of response time adaptation to maximize rewards. In contrast, we hypothesized that a genetic marker of prefrontal dopaminergic function (*COMT*) would be predictive of uncertainty-based exploration and enhanced representation of reward magnitudes.

RESULTS

Temporal integration of expected value

Participants observed a clock arm that completed a revolution over 5 s, and they could stop the clock with a key press in an attempt to win points. Rewards were delivered with a probability and magnitude that varied as a function of response time (RT, **Fig. 1**). The functions were designed such that the expected value (EV; probability × magnitude) increased, decreased or remained constant (IEV, DEV or CEV) with increasing response times (**Fig. 1**). Thus, in the DEV condition, faster RTs yielded more points on average, such that performance benefited from Go learning to produce further speeded RTs. In contrast, fast RTs in the IEV condition yielded below-average outcomes, such that performance benefited from NoGo learning to produce adaptively slower responding. The CEV condition was included for a within-subject baseline RT measure for comparison with IEV and DEV. Because all RTs are equivalently rewarding in the CEV condition, participants' RT in this condition controlled for individual differences in overall motor responding. Given this baseline, an ability to adaptively integrate expected value would be indicated by relatively faster responding in the DEV condition and slower responding in the IEV condition. Dopaminergic manipulations in Parkinson's patients have opposite effects on these measures, likely via modulation of striatal dopamine[29].

We also included a fourth condition (constant expected value–reverse, CEVR) in which reward probability increased while magnitude decreased. This condition serves two purposes. First, because both CEV and CEVR have equal expected values across time, any difference in RT in these two conditions can be attributed to a participants' potential bias to learn more about reward probability than about magnitude or vice versa. Second, CEVR provides another measure of avoidance learning. That is, despite the constant expected value, a bias to learn from negative outcomes will produce slowed responses because of their high probability of occurrence at early response times.

Overall, participants showed robust learning (**Fig. 2**; also see Fig. 5 in the **Supplementary Data Analysis** for RTs for each genotype). Compared to the baseline CEV condition, RTs in the IEV condition were significantly slower ($F_{1,67} = 28.5$, $P < 0.0001$), whereas those in the DEV condition were significantly faster ($F_{1,67} = 6.7$, $P = 0.01$).

There were no effects of any gene either on baseline RTs in the CEV condition or on overall response time (all P values > 0.25). Nevertheless, within-subject RT modulations due to reward structure were predictably altered by striatal genotype (**Fig. 3**). Individuals with the *DARPP-32* T/T genotype showed enhanced Go learning, with faster RTs in the last block of the DEV condition ($F_{1,64} = 4.4$, $P = 0.039$), and, marginally, relative to CEV (DEV_{diff}; $F_{1,64} = 3.1$, $P = 0.08$, an effect that was significant across all trials; $P < 0.05$). *DARPP-32* alleles had no effect on NoGo learning (IEV RTs, or IEV_{diff}; P values > 0.8). Conversely, *DRD2* T/T carriers, who have the highest striatal D2 receptor density[10,15], showed marginally slower RTs in IEV, indicative of enhanced NoGo learning ($F_{1,66} = 3.3$, $P = 0.07$ for both IEV and IEV_{diff}), but no effect on Go learning (P values > 0.3). Modeling results reported below, together with CEVR performance, more strongly support the conclusion that *DARPP-32* and *DRD2* alleles modulate learning to speed and slow RTs from positive and negative outcomes. Finally, there was no effect of *COMT* on any of these measures (P values > 0.35). This constellation of genetic effects converge with those found previously[10] but extend them to a completely different task context, dependent measure and sample. Moreover, these same RT adaptations due to reward structure are sensitive to dopaminergic manipulation in Parkinson's disease[29].

Further analysis revealed genetic contributions to learning from probability relative to magnitude of reinforcement, as assessed by comparing RTs in the CEVR condition (alone and relative to CEV; $P = 0.02$, **Supplementary Data Analysis**). Specifically, individuals

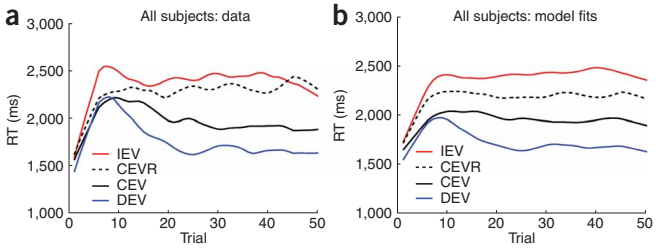

Figure 2 Response times as a function of trial number, smoothed (with weighted linear least-squares fit) over a ten-trial window. (**a**) In all 69 participants. (**b**) Computational model.

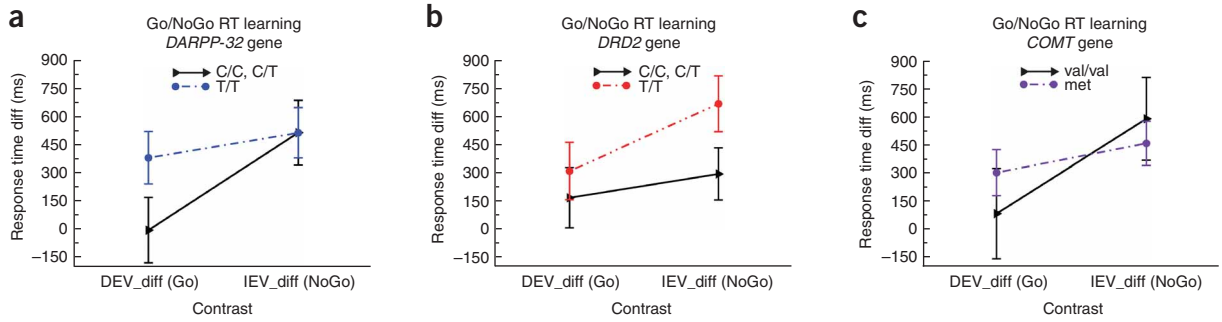

Figure 3 Relative within-subjects biases to speed RTs in DEV relative to CEV (DEV_{diff} = CEV − DEV) and to slow RTs in IEV (IEV_{diff} = IEV − CEV). Values represent mean (s.e.m.) in the last quarter of trials in each condition. (**a**) *DARPP-32* gene. (**b**) *DRD2* gene. (**c**) *COMT* gene.

with enhanced D2 function showed significantly greater sensitivity to frequent negative outcomes in CEVR, again consistent with enhanced NoGo learning. There was also some evidence that carriers of the *COMT* met allele were more sensitive to reward magnitudes (Fig. 1 in **Supplementary Data Analysis**).

Trial-to-trial RT adaptation: exploration?

Although, on average, participants incrementally changed response times dependent on reward structure, single-subject data revealed large RT swings from one trial to the next (**Fig. 4**). These swings did not reflect adaptive changes following rewards or lack thereof[29]. Instead, preliminary analyses indicated that RT swings simply reflected a regression to the mean, whereby faster-than-average responses were more likely to be followed by relatively slower responses and vice versa ($P < 0.0001$; **Supplementary Data Analysis**). As will be clear, however, these RT swings reflect more than just a statistical necessity and are likely to represent participants' tendency to explore the space of responses to determine the reward structure. We investigated this effect in the mathematical reinforcement learning (RL) model developed below.

Computational model

We previously simulated performance in this task using an *a priori* neural network model of the basal ganglia[29]. The model simulates interactive neural dynamics among cortico-striatal circuits and accounts for various effects of dopaminergic manipulation on action selection and reinforcement learning[17,30–32]. Simulated treatments with medications that stimulate dopamine receptors induce speeded RTs in the DEV condition as a result of D1-dependent Go learning in striatonigral cells. However, the same increased dopamine release impedes the ability to slow down in IEV due to excessive D2 receptor stimulation on striatopallidal cells and concomitant impairments in NoGo learning. Simulated dopamine depletion produces the opposite result: less speeding in DEV but better slowing in IEV and CEVR, mirroring the performance of Parkinson's patients on the task[29].

Here we develop an abstract mathematical model designed to quantitatively fit individual participants' response times on a trial-to-trial basis. The purpose of this modeling is threefold: (i) to demonstrate the core computational principles by which the more complex neural model captures the incremental RT changes as a function of reward prediction error; (ii) to augment the model to capture strategic exploratory behavior as a function of reward uncertainty; and (iii) to determine whether best-fitting model parameters for both exploitative and exploratory decisions are predictably modulated as a function of genotype[10].

The point of departure for the model is the central assumption common with virtually all reinforcement models, namely that participants develop an expected value $V(t)$ for the reward they expect to gain in a given trial t. This value is updated as a function of each reward experience using a simple delta rule:

$$V(t+1) = V(t) + \alpha\delta(t)$$

where α is a learning rate that modifies the extent to which values are updated from one trial to the next and δ is the reward prediction error reported by dopamine neurons[16,33], which is simply the reward outcome (Rew) minus the prior expected value:

$$\delta(t) = \mathrm{Rew}(t) - V(t)$$

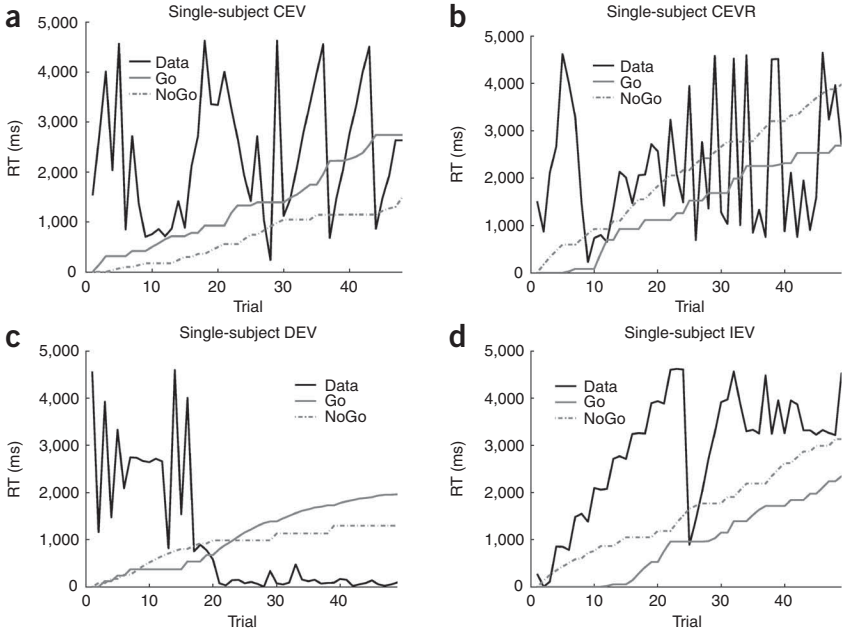

Figure 4 Trial-to-trial RT adjustments in a single subject. (**a**–**d**) Shown are data in CEV (**a**), CEVR (**b**), DEV (**c**) and IEV (**d**). Model Go and NoGo terms (magnified by four times) accumulate as a function of positive and negative prediction errors. Go dominates over NoGo in DEV and the reverse in IEV, but these incremental changes do not capture trial-by-trial dynamics. For this subject, $\alpha_G = 0.63$ and $\alpha_L = 0.74$ (ms per point).

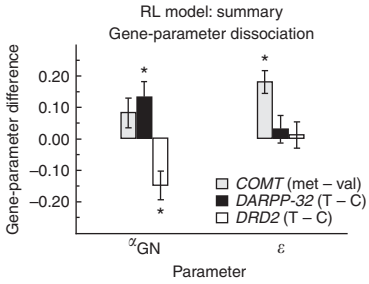

Figure 5 Genetic effects on reinforcement model parameters. *DARPP-32* T/T carriers showed relatively greater learning rates from gains than losses ($\alpha_{GN} = \alpha_G - \alpha_N$) compared to C carriers. *DRD2* T/T carriers showed the opposite pattern. The *COMT* gene did not affect learning rates, but met carriers had significantly higher uncertainty-based explore parameter (ε) values (which are divided by 10^4 to allow them to be displayed on the same scale) than did val/val participants. Error bars, s.e.m.

This value integration is posited to be computed by brain areas upstream of dopamine neurons comprising the "critic," which learns as a function of prediction errors to faithfully represent expected value[5,34,35]. Our model further shares the assumption that these same prediction error signals train the "actor" in the striatum[34]. This process can occur in at least two ways. First, we model a simple, likely implicit, process whereby accumulated positive prediction errors translate into approach-related speeded responses (Go learning), whereas accumulated negative prediction errors produce relative avoidance and slowed responses (NoGo learning)[29,32]. These processes are posited to rely on D1 and D2 receptor mechanisms in separate populations of striatonigral and striatopallidal cells[17,29,32,36]. Because of these differential learning mechanisms, we use different learning rates, and for each:

$$\mathrm{Go}(s,a,t+1) = \mathrm{Go}(s,a,t) + \alpha_G \delta_+(t)$$
$$\mathrm{NoGo}(s,a,t+1) = \mathrm{NoGo}(s,a,t) + \alpha_N \delta_-(t)$$

where α_G controls D1-dependent speeding from positive prediction errors (δ_+) and α_N controls D2-dependent slowing from negative prediction errors (δ_-), for action a and clock-face state s. On each trial RTs were predicted to speed or slow according to differences between current Go and NoGo values.

In addition to this implicit process capturing putative striatal contributions to approach/avoidance, we also model a more strategic process in which participants separately keep track of reward structure for different ('fast' and 'slow') responses (**Supplementary Data Analysis**). With these action representations, participants need only adapt RTs in proportion to the difference between their expected reward values. This would allow, for example, participants to delay responding when slow RTs yield larger rewards on average (as in IEV) or to speed up when they do not. We model this process using Bayesian integration, assuming subjects represent the prior distributions of reward prediction errors separately for fast and slow responses and update them as a function of experience via Bayes' rule:

$$P(\theta|\delta_1 \ldots \delta_n) \propto P(\delta_1 \ldots \delta_n|\theta)P(\theta)$$

where θ reflects the parameters governing the belief distribution about the reward prediction errors for each response, and $\delta_1 \ldots \delta_n$ are the prediction errors observed thus far (on trials 1 to n). Simply stated, Bayes' rule implies that the degree to which each outcome modifies participants' beliefs about obtainable rewards depends on their prior experience and, given this prior, the likelihood that the outcome would occur. As experience is gathered, the means of the posterior distributions accurately represent reward structure in each condition (see below).

We considered that participants either track the probability of a reward prediction error (that is, the probability that a dopamine burst occurs) using beta distributions beta(η,β) or track the magnitude of expected rewards represented by normal distributions $N(\mu,\sigma^2)$. We focus here on the beta distribution implementation, which provided a better fit to the behavioral data. Nevertheless, all genetic results presented below held when using normal distributions and a Kalman filter (**Supplementary Data Analysis**). In either case, RTs were predicted to adapt in proportion to the difference between the best estimates of reward structure for fast and slow responses; that is, the following term was added to the RT prediction: $\rho[\mu_{slow}(s,t) - \mu_{fast}(s,t)]$, where ρ is a free parameter.

We also modeled other parameters that contribute to RT in this task, including simple baseline response speed (irrespective of reward), captured by free parameter K; autocorrelation between the current and previous RT (λ) regardless of reward; and a tendency to adapt RTs toward the single largest reward experienced thus far ('going for gold', parameter υ). Finally, we posited that exploratory strategies would

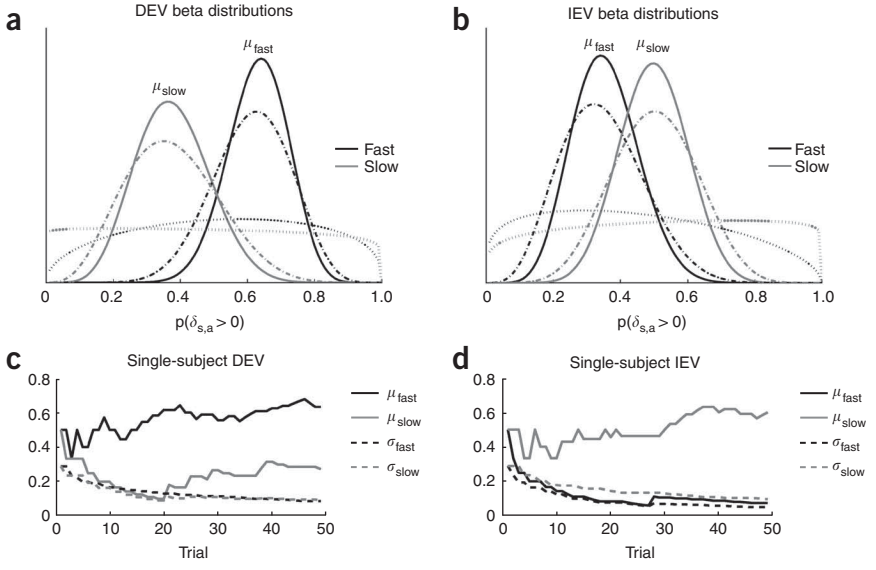

Figure 6 Evolution of action-value distributions. (**a**,**b**) Beta probability density distributions representing the belief about the likelihood of reward prediction errors following fast and slow responses, averaged across all subjects' data. The *x* axis shows the probability of a positive prediction error and the *y* axis represents the belief in each probability, with the mean value μ representing the best guess. Dotted lines reflect distributions after a single trial; dashed lines, after 25 trials; solid lines, after 50 trials. (See **Supplementary Video 1** for dynamic changes in these distributions across all trials for a single subject.). Differences between μ_{fast} and μ_{slow} were used to adjust RTs to maximize reward likelihood. The s.d. σ was taken as an index of uncertainty. Exploration was predicted to modulate RT in direction of greater uncertainty about whether outcomes might be better than the status quo. (**c**,**d**) Trajectory of means and s.d. for a single subject in DEV and IEV conditions. Uncertainties σ decrease with experience. Corresponding beta hyperparameters η and β are shown in **Supplementary Data Analysis**.

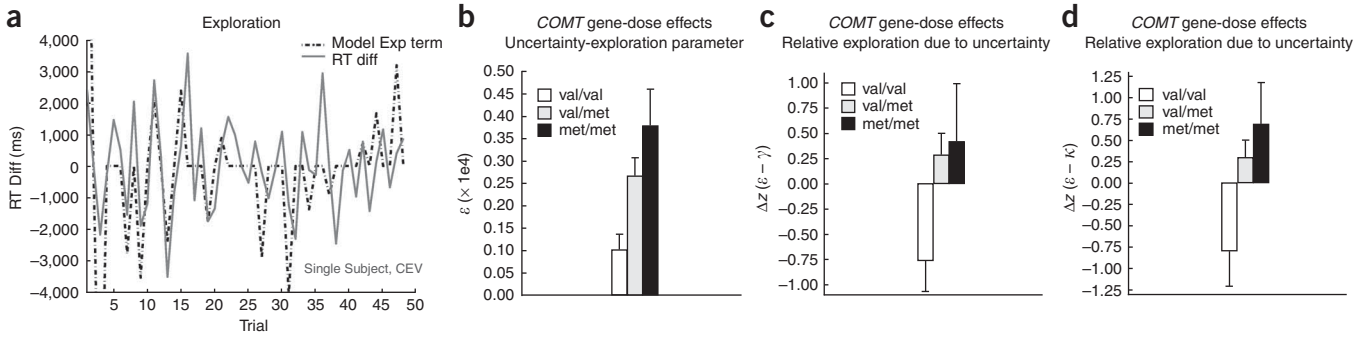

Figure 7 *COMT* gene predicts directed exploration toward uncertain responses. (**a**) RT swings (change in RT from the previous trial) in a single met/met subject in the CEV condition and the corresponding model uncertainty-based Explore term (amplified to be on the same RT scale). See **Supplementary Video 2** for this subject's evolution of beta distributions in CEV. (**b**) Effect of *COMT* gene dose on the uncertainty-based exploration parameter ε. (**c**,**d**) Gene-dose effects were also observed when comparing relative contributions of ε compared with a reverse-momentum parameter γ (**c**) and a lose-switch parameter κ (**d**). Relative *z* scores are plotted here to permit comparison of parameter scaling quantities of different magnitudes. Error bars, s.e.m.

contribute to participants' RT adjustments, as participants sampled the outcomes available to determine which response was most adaptive. This process is modeled as a dynamic Explore process depending on Bayesian uncertainty, which is elaborated further below and is hypothesized to rely on prefrontal cortex–dependent processes. The complete RT update is thus as follows:

$$\hat{\mathrm{RT}}(s,t) = K + \lambda \mathrm{RT}(s,t-1) - \mathrm{Go}(s,a,t) + \mathrm{NoGo}(s,a,t) + \rho[\mu_{\mathrm{slow}}(s,t) - \mu_{\mathrm{fast}}(s,t)] + \nu[\mathrm{RT}_{\mathrm{best}} - \mathrm{RT}_{\mathrm{avg}}] + \mathrm{Explore}(s,t)$$

For each subject, a single set of best fitting parameters was derived across all conditions. The model captures the qualitative pattern of results, with predicted RT changing as a function of reward structure (**Fig. 2b**; see Fig. 6 in **Supplementary Data Analysis** for model fits for each genotype). Positive prediction errors are most prevalent for early responses in DEV, and accordingly model RTs are fastest in this condition. Negative prediction errors are most prevalent in IEV and CEVR, leading to slowed model responses.

We hypothesized that these relative learning rate parameters for determining exploitative responses would be modulated by striatal genotype. Indeed, *DARPP-32* T/T carriers, who should have increased striatal D1-dependent learning[10,13,14], had relatively larger α_G as compared to α_N than did C carriers, suggesting relatively greater sensitivity to positive than negative prediction errors (**Fig. 5**; $F_{1,65} = 4.0$, $P = 0.05$). Conversely, *DRD2* T/T carriers, with relatively greater D2 receptor density[29], showed relatively greater learning from negative prediction errors ($F_{1,66} = 5.3$, $P = 0.02$). Relative learning rates were not modulated by *COMT* genotype ($P > 0.2$), and other than the Explore parameter, no other parameters differed as a function of any genotype (all P values > 0.2).

Uncertainty-based exploration

The above model provides an account of incremental RT changes as a function of reward prediction error, and it provides evidence for the mechanisms posited to mediate these effects in neural networks[29]. Nevertheless, inspection of individual subject data reveals more complex dynamics than those observed in the averaged data (**Fig. 4**). These plots show RTs across trials for an arbitrary single participant, along with model Go and NoGo terms. Asymptotically, the participant converges on a faster RT in DEV, and slower RT in IEV, relative to CEV. However, at the more fine-grained scale, there are often large RT swings from one trial to the next that are not captured by model learning mechanisms.

We hypothesized that these RT swings are rational, in that they might reflect exploratory strategies to gather statistics of reward structure. Several solutions have been proposed to manage the exploration/exploitation tradeoff. If performance is unsatisfactory over extended periods, stochastic noise can simply be added to behavioral outputs, promoting random exploratory choices[7]. Alternatively, exploration can be strategically directed toward particular choices in proportion to the amount of information that would be gained, regardless of past performance[4,6,37,38]. Our model embodies the assumption that exploratory decisions occur in proportion to the participant's relative uncertainty about whether responses other than those currently being exploited might yield better outcomes. This assumption builds on prior modeling in which exploration is encouraged by adding an 'uncertainty bonus' to the value of decision options having uncertain outcomes[4,6,37,38]. Here we posit that exploration occurs in proportion to uncertainty about the probability that the explored option will yield a positive reward prediction error (or, in alternative models, uncertainty about the expected value of such rewards or reward prediction errors; **Supplementary Data Analysis**). The Bayesian framework for integrating reward statistics provides a natural index of uncertainty: the s.d. of the prior distributions[39], which decreases after sampling a given action (albeit at a slower rate for more variable outcomes).

Initially, distributions representing belief about reward structure for each response category are wide, reflecting maximum uncertainty (**Fig. 6**). As experience with each option is gathered, the distributions evolve to reflect the underlying reward structures, such that the mean belief is higher for fast responses in DEV and for slow responses in IEV. Moreover, the s.d., and hence uncertainties, decrease with experience. This process is analogous to estimating the odds of a coin flip resulting in heads or tails, with uncertainty about those odds decreasing with the number of observations. With these distributions, the relative uncertainties for fast and slow responses in a given trial can be used as a rational heuristic to drive exploration. In particular, the Explore term of the model is computed as follows:

$$\mathrm{Explore}(s,t) = \varepsilon[\sigma_{\sigma|s,a=\mathrm{Slow}} - \sigma_{\sigma|s,a=\mathrm{Fast}}]$$

where ε is a free parameter that scales exploration in proportion to relative uncertainty and $\sigma_{\delta|s,a=\mathrm{Slow}}$ and $\sigma_{\delta|s,a=\mathrm{Fast}}$ are the standard

deviations quantifying uncertainty about reward prediction error likelihood given slow and fast responses, respectively. Thus, with sufficiently high ε, RT swings are predicted to occur in the direction of greater uncertainty about the likelihood that outcomes might be better than the status quo.

Overall, including this uncertainty-based exploration term provided a better fit to trial-by-trial choice than the base model without exploration (and penalizing the fit for the additional parameters; see **Supplementary Data Analysis**). Although the model cannot deterministically predict RT swings (which reflect the output of multiple interacting processes, including those sensitive to previous reinforcement), there is nevertheless a reliable positive correlation between the model's uncertainty-based exploratory predictions and participants' actual RT swings from on trial to the next ($r_{4,214} = 0.31$, $P < 0.0001$; **Fig. 7**; Fig. 3 in **Supplementary Data Analysis**).

Moreover, this relationship was particularly evident for carriers of the *COMT* met allele (Fig. 3 in **Supplementary Data Analysis**), supporting a role for PFC neuromodulatory control over exploration as a function of decision uncertainty. The ε parameter that scales exploration in proportion to uncertainty was significantly higher among met allele carriers (**Fig. 5**; $F_{1,67} = 8.2$, $P = 0.006$). Further, there was a monotonic gene-dose effect, with ε values being largest in met/met participants, intermediate in val/met and smallest in val/val carriers (**Fig. 7b**; $F_{1,67} = 9.5$, $P = 0.003$). No such effects on ε were observed for *DARPP-32* or *DRD2* genotypes (P values > 0.5).

Notably, the *COMT* exploration effects appear to be specific to uncertainty. First, overall RT variability (in terms of s.d.) did not differ as a function of genotype ($P > 0.2$). Second, a number of foil models attempting to account for RT swings without recourse to uncertainty confirmed that only the uncertainty-based exploration parameter can account for *COMT* effects (**Supplementary Data Analysis**). For example, we included a 'reverse momentum' parameter γ, which predicted RT swings to counter a string of progressively speeded or slowed responses, regardless of uncertainty. Although this model provided a reasonable fit to RT swings overall, the uncertainty model was superior only in carriers of the *COMT* met allele (**Supplementary Data Analysis**). We also included a 'lose-switch' parameter κ, which predicted RTs to adjust from fast to slow or vice versa following a negative prediction error. Notably, there were *COMT* gene-dose effects not only on raw ε values but also on their relative weighting compared to either γ or κ (P values < 0.004; **Fig. 7c,d**). This result implies that the contribution of *COMT* to RT swings is specific to uncertainty.

DISCUSSION

Individuals differ substantially in their motivational drives. The present findings demonstrate three distinct aspects of value-based decision-making associated with independent genetic factors (see summary **Fig. 5**). These genes modulate specific aspects of dopaminergic function in brain areas thought to support exploration and exploitation[6,7,10,19]. Behaviorally, exploitative choices were manifest by RT differences between conditions in which rewards could on average be maximized by responding earlier (DEV) or later (IEV) in the trial, compared to baseline (CEV) conditions. Modeling showed that striatal genetic effects are accounted for by individual differences in learning rates from positive and negative prediction errors and their coupling with response speeding and slowing. This result is nontrivial: striatal genes could have affected exploitation by modulating the extent to which RTs are adjusted as a function of mean reward value estimates (that is, the ρ parameter). Similarly, whereas trial-to-trial RT swings were readily viewable in single-subject data (**Fig. 4**), the specific components due to uncertainty-based exploration, and individual differences therein, were only extracted with the computational analysis.

Our observation that *DARPP-32* and *DRD2* modulate reinforcement learning in the temporal decision-making domain is consistent with similar genetic effects in choice paradigms[10] and with data from patients with Parkinson's disease, on and off medication, in this same task[29]. Recent rodent studies show direct support for the model's dual D1 and D2 mechanisms of synaptic plasticity[17,18].

The present human genetic data provide support for the mechanisms posited in models of striatal dopamine, in which accumulated reward prediction errors over multiple trials produce speeded responses, whereas negative prediction errors slow responses[29,40]. Our assumption that *DARPP-32* genetic effects reflect striatal D1 receptor–mediated Go learning is supported by evidence that the DARPP-32 protein is highly concentrated in the striatum[12] and is critical for D1- but not D2-dependent synaptic plasticity and behavioral reward learning[13,14]. These data also converge with effects of pharmacological manipulation of striatal D1 receptors on appetitive approach and response speeding to obtain rewards in monkeys and rats[36,41].

Similarly, our assumption that *DRD2* genetic effects reflect primarily striatal D2 receptor–mediated learning is supported by evidence that T/T carriers show enhanced striatal D2 receptor density[29,42]. Theoretically, striatal D2 receptors are thought to be necessary for learning in striatopallidal neurons when dopamine levels are low[17], as is the case during negative prediction errors[43–45] or as a result of Parkinson's disease[29,30]. Indeed, synaptic potentiation in striatopallidal neurons is elevated under conditions of dopamine depletion[18]. Conversely, rats with reduced striatal D2 receptor density[46] are less sensitive to aversive outcomes, persisting in taking addictive drugs even when this is followed by shocks[47].

Perhaps less clear is the precise neurobiological mechanism by which *COMT* modulates uncertainty-based exploration. Indeed, the mechanisms of exploration are understudied compared to those of exploitation. Nevertheless, neuroimaging studies reveal that in non-reinforcement-learning contexts, anterior prefrontal cortical regions reflect Bayesian uncertainty[21], and that this same region is activated when participants make exploratory decisions in a RL environment[6]. Our findings provide the first evidence for exploratory decisions that occur in proportion to uncertainty about whether other responses might produce better outcomes than the status quo. This exploration strategy is strongly motivated by prior theoretical work[6,7,38] and seems to be highly dependent on genetic function in the prefrontal cortex. Furthermore, the *COMT* effects on trial-to-trial 'lose-shift' behavior in choice paradigms that we originally reported[10] might be more parsimoniously explained by uncertainty-based exploratory mechanisms. Indeed, in that study, met carriers showed greater propensity to shift only in the initial trials of the task, when reward structure was most uncertain. Thus, these exploratory strategies may be viewed as an attempt to minimize uncertainty.

In contrast to the multiple extant neural models of exploitation, there is a dearth of models investigating how neuronal populations can learn to represent quantities of uncertainty as a function of experience. Nevertheless, the sorts of Bayesian probability distributions required for the uncertainty computations used here are naturally coded in populations of spiking neurons[48,49]. Thus, future research should examine how such representations can be learned and whether prefrontal dopamine supports the uncertainty computations *per se*, the active maintenance of relative uncertainties in working memory across trials, or simply the final decision to override exploitative strategies in order to explore when uncertainty is sufficiently high.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank S. Williamson and E. Carter for help with DNA analysis and administering cognitive tasks to participants, and N. Daw, P. Dayan, and R. O'Reilly for helpful discussions. This research was supported by US National Institutes of Mental Health grant R01 MH080066-01.

AUTHOR CONTRIBUTIONS

M.J.F., B.B.D. and F.M. designed the study; M.J.F. conducted the modeling and analyzed the behavioral data; B.B.D. collected data; J.O.-T. and F.M. extracted the DNA and conducted genotyping; M.J.F., B.B.D. and F.M. wrote the manuscript.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprintsandpermissions/.

1. Scheres, A. & Sanfey, A.G. Individual differences in decision making: Drive and Reward Responsiveness affect strategic bargaining in economic games. *Behav. Brain Funct.* **2**, 35 (2006).
2. Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C.F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680–1683 (2005).
3. Frank, M.J., Woroch, B.S. & Curran, T. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron* **47**, 495–501 (2005).
4. Gittins, J.C. & Jones, D. A dynamic allocation index for the sequential design of experiments. in *Progress in Statistics* (eds. Gani, J., Sarkadi, K. & Vincze, I.), 241–266 (North Holland Publishing Company, Amsterdam, 1974).
5. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, USA, 1998).
6. Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
7. Cohen, J.D., McClure, S.M. & Yu, A.J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. R. Soc. Lond. B* **362**, 933–942 (2007).
8. Depue, R.A. & Collins, P.F. Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation, and extraversion. *Behav. Brain Sci.* **22**, 491–517 (2001).
9. Meyer-Lindenberg, A. *et al.* Genetic evidence implicating DARPP-32 in human frontostriatal structure, function, and cognition. *J. Clin. Invest.* **117**, 672–682 (2007).
10. Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T. & Hutchison, K.E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl. Acad. Sci. USA* **104**, 16311–16316 (2007).
11. Klein, T.A. *et al.* Genetically determined differences in learning from errors. *Science* **318**, 1642–1645 (2007).
12. Ouimet, C.C., Miller, P.E., Hemmings, H.C., Walaas, S.I. & Greengard, P. DARPP-32, a dopamine- and adenosine 3':5'-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. III. Immunocytochemical localization. *J. Neurosci.* **4**, 111–124 (1984).
13. Stipanovich, A. *et al.* A phosphatase cascade by which rewarding stimuli control nucleosomal response. *Nature* **453**, 879–884 (2008).
14. Calabresi, P. *et al.* 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 (2000).
15. Hirvonen, M. *et al.* Erratum: C957T polymorphism of the dopamine D2 receptor (*DRD2*) gene affects striatal DRD2 availability *in vivo*. *Mol. Psychiatry*. **10**, 889 (2005).
16. Montague, P.R., Dayan, P. & Sejnowski, T.J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
17. Frank, M.J. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J. Cogn. Neurosci.* **17**, 51–72 (2005).
18. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D.J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848–851 (2008).
19. Graybiel, A.M. Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* **31**, 359–387 (2008).
20. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).
21. Yoshida, W. & Ishii, S. Resolution of uncertainty in prefrontal cortex. *Neuron* **50**, 781–789 (2006).
22. Frank, M.J. & Claus, E.D. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Rev.* **113**, 300–326 (2006).
23. Roesch, M.R. & Olson, C.R. Neuronal activity related to reward value and motivation in primate frontal cortex. *Science* **304**, 307–310 (2004).
24. Rudebeck, P.H., Walton, M.E., Smyth, A.N., Bannerman, D.M. & Rushworth, M.F.S. Separate neural pathways process different decision costs. *Nat. Neurosci.* **9**, 1161–1168 (2006).
25. Meyer-Lindenberg, A. *et al.* Midbrain dopamine and prefrontal function in humans: interaction and modulation by COMT genotype. *Nat. Neurosci.* **8**, 594–596 (2005).
26. Slifstein, M. *et al.* COMT genotype predicts cortical-limbic D1 receptor availability measured with [^{11}C]NNC112 and PET. *Mol. Psychiatry* **13**, 821–827 (2008).
27. Gogos, J.A. *et al.* Catechol-O-methyltransferase-deficient mice exhibit sexually dimorphic changes in catecholamine levels and behavior. *Proc. Natl. Acad. Sci. USA* **95**, 9991–9996 (1998).
28. Forbes, E.E. *et al.* Genetic variation in components of dopamine neurotransmission impacts ventral striatal reactivity associated with impulsivity. *Mol. Psychiatry* **14**, 60–70 (2009).
29. Moustafa, A.A., Cohen, M.X., Sherman, S.J. & Frank, M.J. A role for dopamine in temporal decision making and reward maximization in parkinsonism. *J. Neurosci.* **28**, 12294–12304 (2008).
30. Frank, M.J., Seeberger, L.C. & O'Reilly, R.C. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).
31. Santesso, D., Evins, A., Frank, M., Cowman, E. & Pizzagalli, D. Single dose of a dopamine agonist impairs reinforcement learning in humans: evidence from event-related potentials and computational modeling of striatal-cortical function. *Hum. Brain Mapp.* **30**, 1963–1976 (2009).
32. Wiecki, T.V., Riedinger, K., Meyerhofer, A., Schmidt, W.J. & Frank, M.J. A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: context dependency, extinction, and renewal. *Psychopharmacology (Berl.)* **204**, 265–277 (2009).
33. Bayer, H.M. & Glimcher, P.W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
34. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
35. O'Reilly, R.C., Frank, M.J., Hazy, T.E. & Watz, B. PVLV: the primary value and learned value Pavlovian learning algorithm. *Behav. Neurosci.* **121**, 31–49 (2007).
36. Nakamura, K. & Hikosaka, O. Role of dopamine in the primate caudate nucleus in reward modulation of saccades. *J. Neurosci.* **26**, 5360–5369 (2006).
37. Sutton, R.S. Integrated architectures for learning, planning and reacting based on approximating dynamic programming. *Proceedings of the Seventh International Conference on Machine Learning* (Porter, B.W. & Mooney, R.J., eds.) 216–224 (Morgan Kaufmann, Palo Alto, California, USA, 1990).
38. Dayan, P. & Sejnowski, T.J. Exploration bonuses and dual control. *Mach. Learn.* **25**, 5–22 (1996).
39. Daw, N.D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
40. Niv, Y., Daw, N.D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl.)* **191**, 507–520 (2007).
41. Dalley, J.W. *et al.* Time-limited modulation of appetitive Pavlovian memory by D1 and NMDA receptors in the nucleus accumbens. *Proc. Natl. Acad. Sci. USA* **102**, 6189–6194 (2005).
42. Zhang, Y. *et al.* Polymorphisms in human dopamine D2 receptor gene affect gene expression, splicing, and neuronal activity during working memory. *Proc. Natl. Acad. Sci. USA* **104**, 20552–20557 (2007).
43. Hollerman, J.R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
44. Satoh, T., Nakai, S., Sato, T. & Kimura, M. Correlated coding of motivation and outcome of decision by dopamine neurons. *J. Neurosci.* **23**, 9913–9923 (2003).
45. Bayer, H.M., Lau, B. & Glimcher, P.W. Statistics of midbrain dopamine neuron spike trains in the awake primate. *J. Neurophysiol.* **98**, 1428–1439 (2007).
46. Dalley, J.W. *et al.* Nucleus accumbens D2/3 receptors predict trait impulsivity and cocaine reinforcement. *Science* **315**, 1267–1270 (2007).
47. Belin, D., Mar, A.C., Dalley, J.W., Robbins, T.W. & Everitt, B.J. High impulsivity predicts the switch to compulsive cocaine-taking. *Science* **320**, 1352–1355 (2008).
48. Zemel, R.S., Dayan, P. & Pouget, A. Probabilistic interpretation of population codes. *Neural Comput.* **10**, 403–430 (1998).
49. Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).

ONLINE METHODS

Sample. We tested 73 healthy participants who were recruited from the University of Arizona undergraduate psychology subject pool and who provided informed written consent. Two subjects declined genetic sampling and are excluded from analysis. Failed genetic assays eliminated a further two *COMT* samples, two *DRD2* samples and three *DARPP-32* samples. The remaining 69 subjects (46 female) had a mean age of 19 (s.e.m. = 0.2) and comprised 48 self-identified as Caucasian, 14 Hispanics, 2 Asians, 1 African-American and 4 subjects who categorized themselves as 'Other'. The breakdown of *COMT* genotypes was 19:43:7 (val/val:val/met:met/met). The breakdown of *DRD2* genotypes was 31:38 (C carriers:T/T homozygotes). The breakdown of *DARPP-32* genotypes was 38:29 (T/T:C carriers; note that in our prior report the T/T genotype was incorrectly referred to as A/A, and C carriers as G carriers, due to mislabeling of the base-pair complement[10]. Thus, the T/T subjects here reflect the same genotype previously associated with enhanced Go learning.) Genetic effects were independent: there was no association between the distribution of any one polymorphism and any other (for example, *DRD2* genotype was not predictive of *COMT* genotype, and so on Fisher's exact test, $P > 0.3$). All genotypes were in Hardy-Weinberg equilibrium (P values > 0.1), with the exception of *COMT* ($\chi^2[1] = 5.6$, $P < 0.05$). This deviation is likely to be due to heterogeneity in the population; in an analysis of individuals self-identifying as Caucasian alone, Hardy-Weinberg equilibrium was not violated ($P > 0.1$).

Genotyping. Genotyping procedures were carried out in the Molecular Psychiatry Laboratory at the University of Arizona. DNA samples were extracted from saliva samples using Oragene DNA Collection Kits (DNAGenotek). Genomic DNA was amplified using standard PCR protocols.

Dopamine- and adenosine-3′,5′-monophosphate (cAMP)-regulating phosphoprotein SNP (encoded by *DARPP-32*, rs907094). Genomic DNA was amplified for the *DARPP-32* (also called *PPP1R1B*) SNP using standard PCR protocols. Amplification of the 404-bp region was carried out using the sense primers DD-F 5′-GCATTGCTGAGTCTCACCTGCAGTCT-3′ and antisense primers DD-R 5′-ATTGGGAGAGGGACTGAGCCAAGGATGG-3′ in a reaction volume of 25 µl consisting of 2.5 ng of DNA, 0.25 mM dNTPs, 0.25 µM each sense and antisense primers, 1× Qiagen PCR buffer and 1.5 U *Taq* DNA polymerase (Qiagen). Thermocycling conditions consisted of an initial denaturation step of 95 °C for 5 min followed by 35 cycles of 94 °C for 30 s, 72 °C for 60 s, and 72 °C for 60 s, with a final extension step of 72 °C for 10 min. PCR products were sequenced using the ABI 3730XL DNA Analyzer (Applied Biosystems) and visualized using Chromas Vs. 2.13 (Technelysium).

***COMT* rs4680.** Genomic DNA was amplified for the *Comt4680* polymorphism using standard PCR protocols. Amplification of the 109-bp region was carried out using the sense primers Comt-F 5′-TCTCCACCTGTGCTCACCTC-3′ and antisense primers Comt-R 5′-GATGACCCTGGTGATAGTGG-3′ in a reaction volume of 25 µl consisting of 2.5 ng of DNA, 0.25 mM dNTPs, 0.25 µM each sense and antisense primers, 1× Qiagen PCR buffer and 1 U *Taq* DNA polymerase (Qiagen). Thermocycling conditions consisted of an initial denaturation step of 95 °C for 5 min followed by 35 cycles of 95 °C for 15 s, 54 °C for 20 s, and 72 °C for 30 s, with a final extension step of 72 °C for 5 min. The restriction enzyme NlaIII (5 U, New England Biolabs) was added to a 20-µl aliquot of the PCR product and digested for 2 h at 37 °C. Five microliters of the digested PCR product was added to 4 µl of Orange G DNA loading buffer and loaded onto a 3% agarose gel. Images were captured via the Gel Doc XR System (Bio-Rad).

***DRD2* rs6277.** Optimization of tetra-primer ARMS PCR for the detection of the DRD2 polymorphism was performed empirically using primers designed by original software developed by the founders of the tetra-primer ARMS PCR method and available on the website http://cedar.genetics.soton.ac.uk/public_html/primer1.html, with a T_m optimized to 72 °C and a GC content of 48.7%.

Genomic DNA was amplified for the *DRD2* polymorphism using tetra-primer ARMS PCR protocol as described[50]. Amplification of the total 2,950-bp region was carried out using the outer sense primers DRD2-F 5′-ACGGCTC ATGGTCTTGAGGGAGGTCCGG-3′ and outer antisense primers DRD-R 5′-CCAGAGCCCTCTGCCTCTGGTGCAGGAG-3′ as well as inner sense primers DRD-Fi 5′-ATTCTTCTCTGGTTTGGCGGGGCTGGCA-3′ and inner antisense primers 5′-CGTCCCACCACGGTCTCCACAGCACTACC-3′ in a reaction volume of 25 µl consisting of 2.5 ng of DNA, 0.25 mM dNTPs, 0.025 µM outer sense and antisense primers, 0.25 µM inner sense and antisense primers, 1× Qiagen PCR buffer and 2 U *Taq* DNA polymerase (Qiagen). Thermocycling conditions consisted of an initial denaturation step of 95 °C for 5 min followed by 35 cycles of 94 °C for 30 s, 72 °C for 60 s, and 72 °C for 60 s, with a final extension step of 72 °C for 10 min. Five microliters of the PCR product was added to 4 µl of Orange G DNA loading buffer and loaded onto a 3% agarose gel and run in 0.5× TAE buffer for 20 min at 72 V. The gels were prestained with GelStar Nucleic Acid Gel Stain and images were captured with the Gel Doc XR System (Bio-Rad).

Genotyping for *DRD2* was carried in triplicate, and identification of each individual allele was conducted by three independent observers with 100% agreement.

Ethnicity. Because there was some heterogeneity in the sample (14 subjects were Hispanic), it was critical to establish whether genetic effects could have been due occult stratification. To this end, we reanalyzed the data with the 14 Hispanic individuals omitted and found very similar patterns of results for each genotype. Similar results also were found when omitting all individuals not self-identifying as Caucasian. We also reanalyzed all the data and included an additional factor into the general linear model according to whether subjects were Hispanic or not. In this analysis, all genetic effects remained significant and there was no effect of ethnicity, nor was there an interaction between ethnicity and genotype (P values > 0.25). Again, similar findings were included if the factor coded whether subjects were self-identifying as Caucasian or non-Caucasian. Finally, Hardy-Weinberg equilibrium data were also analyzed when excluding Hispanic and other individuals not self-identifying as Caucasian, and no genotype frequencies deviated from equilibrium.

Task methods. Task instructions were as follows:

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

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

"*Hint: Try to respond at different times along the clock cycle in order to learn how to make the most points. Note: The length of the experiment is constant and is not affected by when you respond.*" This hint was provided to prevent participants from responding quickly simply to leave the experiment early and in an attempt to equate reward rate (that is, rewards per second) across conditions. In addition, earlier responses were associated with longer intertrial intervals so that the statement that the length of the experiment was constant was roughly accurate. However, because subjects might be averse to waiting through long intertrial intervals, and because we also wished to reduce the predictability of the onset of the next trial's clock face stimulus, we set the intertrial interval to (5,000 – RT)/2. Thus, faster responses were associated with longer wait times, but the onset of each trial was temporally unpredictable.

The order of condition (CEV, DEV, IEV, CEVR) was counterbalanced across participants. A rest break was given between each of the conditions (after every 50 trials). Subjects were instructed at the beginning of each condition to respond at different times in order to try to win the most points but were not told about the different rules (for example, IEV, DEV). Each condition was also associated with a different color of clock face to facilitate encoding that the participant was in a new context, with the assignment of condition to color counterbalanced. Participants completed 50 trials of one condition before proceeding to the next, for a total of 200 trials.

To prevent participants from explicitly memorizing a particular value of reward feedback for a given response time, we also added a small amount of random uniform noise (±5 points) to the reward magnitudes on each trial.

Analysis. General linear models were used for all statistical analysis. *COMT* gene-dose effects were tested by entering the number of met alleles expressed by each subject as a continuous variable. Behavioral analyses, except where indicated, examined RTs in the last quarter (12 trials) of each condition, by

which time participants were likely to have learned the reward structure of the particular clock face[29]. (Although it is possible to compute learning from the first to last quarter of each condition, some participants learned to discriminate reward structure even in the first quarter, minimizing the difference across quarters. We therefore focused our analyses on the last quarter, in which performance was expected to stabilize. Further, the model-based analyses converge with those derived from these behavioral measures without confining analysis to any part of the learning curve.) In some analyses, the degrees of freedom are 1 less than they should be because a computer crash occurred for one subject who therefore did not complete all conditions.

Model methods. In all models, we used the Simplex method with multiple starting points to derive best-fitting parameters for each individual participant that minimized the sum of squared error (SSE) between predicted and actual RTs across all trials. A single set of parameters was derived for each subject providing the best fit across all task conditions. Data were smoothed with a five-trial moving average for fitting of sequential time-series responses, although similar results were produced without such smoothing, just with larger overall SSEs for all models. Model fits were evaluated with Akaike's Information Criterion (AIC), which penalizes model fits for models with additional parameters:

$$\mathrm{AIC} = 2k + n[\log(2\pi \mathrm{SSE}/n) + 1]$$

where k is the number of parameters, n is the number of data points to be fit and SSE is the sum of squared error between the model predictions and actual response times across all trials for each subject. The model with the lowest AIC value is determined to be the best fit.

Exploit model. There are several ways in which RTs might be modeled in this task. Our first aim was to derive a simple model to approximate the mechanisms embodied within our a priori neural network model of the basal ganglia, which predicted the double dissociation between RTs in the DEV and IEV conditions dependent on dopaminergic medication status in Parkinson's disease[29]. Because that model is complex and involves multiple interacting brain areas, we sought to capture its core computations in abstract form and to then fit free parameters of this reduced model to individual subject data, which in turn can be linked to striatal dopaminergic genes. A similar procedure was used in a choice rather than RT task[10].

We modeled the incremental RT changes in the different conditions via separate Go and NoGo parameters that learn from positive and negative prediction errors and serve to speed and slow RTs, respectively. These parameters correspond to D1- and D2-dependent learning in striatonigral and striatopallidal neurons. The terms 'Go' and 'NoGo' are shorthand descriptions of the functions of the two pathways in the neural model, whereby Go and NoGo activity separately report the learned probability that a given action in the current state would produce a positive and negative outcome, respectively. In choice paradigms, the probability that an action is taken is proportional to the relative (Go – NoGo) activity for that action, as compared to all other actions. Here, as far as the striatum is concerned in the model, there is only one action ("hit the spacebar"), and the relative (Go – NoGo) activity simply determines the speed at which that action is executed.

Positive and negative prediction errors are computed relative to current expected value V, which are then used to update V estimates for subsequent trials and also to train the Go and NoGo striatal values. This scheme is reminiscent of "actor-critic" reinforcement learning models[5,34], where the critic is the V system, the prediction errors of which are reflected in phasic dopaminergic signals, and the actor comprises Go and NoGo striatal neuronal populations[17,29].

The expected value V was initialized to 0 at the beginning of the task. The final V value at the end of each condition was carried over to the beginning of the next, on the assumption that any rewards obtained at the beginning of a condition are compared relative to their best estimate of expected value in the task at large (for example, 50 points might be interpreted as a positive prediction error if in the last block they had on average obtained 20 points, but would be a negative prediction error if their previous average point value was 100). Go and NoGo values were initialized to 0 and accumulated as a function of reward prediction errors for each state (clock face). (Although the Go and NoGo terms accumulate monotonically as a function of experience, in the neural model, Go synapses are weakened following negative prediction errors and NoGo synapses are strengthened, preventing these values from saturating. Here the contributions of Go and NoGo terms were small enough for this to not be necessary; however, adding a decay term to Go/NoGo values to prevent increases without bound did not change the basic pattern of results.) Finally, due to model degeneracy, α was held constant and was set to 0.1 to allow integration of history, allowing other Go/NoGo learning parameters to vary freely. This same critic learning rate was used in the neural network implementation[29].

Bayesian integration of expected value. The Go and NoGo learning mechanisms capture a relatively automatic process in which the striatum speeds or slows responses after positive or negative prediction errors, respectively, independent of the RTs that produced those reinforcements. This mechanism may result from the architecture of the basal ganglia, which supports approach and avoidance behavior for positive and negative outcomes. This mechanism is also adaptive in the current task if participants' initial responses are faster than the midpoint (as was typically the case), in which case positive prediction errors predominate in DEV and negative prediction errors predominate in IEV, leading to speeding and slowing, respectively. The improved behavioral fit (including penalty for additional parameters) provided by including these mechanisms suggests that these tendencies capture some of the variance in this task. However, note that these mechanisms are not necessarily adaptive in all cases: for example, slow responses that produce positive prediction errors (for example, in IEV) would lead to subsequent speeding according to this mechanism.

We posited that in addition to Go/NoGo learning, subjects would attempt to explicitly keep track of the rewards experienced for different responses and then produce those responses that had been rewarded most. It is unrealistic to assume that participants track reward structure for all possible response times. Instead, we employed a simplifying (and perhaps more plausible) assumption that participants simply track reward structure for responses categorized as "fast" or "slow." Given that the reward functions are monotonic (and assuming subjects believe this to be the case), one only needs to track rewards separately for fast and slow responses to determine which has the highest expected value, and to respond faster or slower in proportion to the difference in these values.

We thus categorized each response depending on whether it was faster or slower than the participants local mean $\mathrm{RT}_{\mathrm{avg}}$, which was itself tracked with the delta rule:

$$\mathrm{RT}_{\mathrm{avg}}(t) = \mathrm{RT}_{\mathrm{avg}}(t-1) + \alpha[\mathrm{RT}(t-1) - \mathrm{RT}_{\mathrm{avg}}(\mathrm{t}-1)]$$

(This choice for tracking average RT was not critical; all results are similar even if simply defining fast and slow according to the first and second halves of the clock. However, using an adaptive local mean RT is more general and may prove useful if the reward functions are nonmonotonic.)

We represented participants' beliefs about reward structure for these two response categories in Bayesian terms, assuming participants represent not only a single value of each response but rather a distribution of such values and, crucially, the uncertainty about them[39]. In particular, we posited that participants would track the estimated likelihood of obtaining a positive reward prediction error for each response, or the magnitude of such prediction errors, as a function of the past set of dopamine bursts reported by midbrain dopamine neurons. Any probability distribution in the exponential family of distributions can be represented in a population of spiking neurons[48,49], so a priori it is not clear whether it is more plausible for participants to track simply the probability of a dopamine burst occurring at all or to instead represent the magnitude of the typical prediction error. Model fits to data were clearly superior for probability simulations, which we focus on here; nevertheless, as reported in the **Supplementary Data Analysis**, all genetic findings hold when modeling reward magnitudes (or reward prediction error magnitudes) with a Kalman filter.

We represented the likelihood of reward prediction errors for each state s and fast or slow action a as beta distributions $\mathrm{beta}(\eta_{s,a},\beta_{s,a})$ (see below). The probability of a reward prediction error can be represented as a binomial process, and the beta distribution is the conjugate prior to the binomial distribution. This implies that the application of Bayes' rule to update the

doi:10.1038/nn.2342

prior distribution results in a posterior distribution that is itself also a beta distribution with new parameters. (Strictly speaking, a binomial process assumes that each observation is independent. This assumption is violated in the case of reward prediction errors because a given reward value may be interpreted as a positive or negative prediction error depending on prior reinforcement context. The beta distribution is nevertheless a simplifying assumption that provided a substantial improvement to behavioral fit. Furthermore, we also modeled a version in which we tracked the probability of obtaining a nonzero reward, rather than a reward prediction error. In this model, we also binarized responses such that "fast" and "slow" responses were categorized according to those that were in the first and second halves of the clock. In this case, each observation is indeed independent, and all core results continued to hold.)

The probability density function of the beta distribution is as follows:

$$f(x;\eta,\ \beta) = \frac{x^{\eta-1}(1-x)^{\beta-1}}{\int_0^1 z^{\eta-1}(1-z)^{\beta-1}dz}$$

where the integral in the denominator is the beta function $B(\eta,\beta)$ and is a normalization factor that ensures that the area under the density function is always 1. The defining parameters of the posterior distribution for each state s are calculated after each outcome using Bayes' rule:

$$P(\eta,\beta|\delta_1 \ldots \delta_n) = \frac{P(\delta_1 \ldots \delta_n|\eta,\beta)P(\eta,\beta)}{\int\int P(\delta_1 \ldots \delta_n|\eta,\beta)d\eta d\beta} = \frac{P(\delta_1 \ldots \delta_n|\eta,\beta)P(\eta,\beta)}{P(\delta_1 \ldots \delta_n)}$$

Explore model. Because of the conjugate prior relationship between binomial and beta distributions, this update is trivial without having to directly compute Bayes' equation above. The η and β parameters are updated for each state or action by simply incrementing the prior η and β hyperparameters after each instance of a positive or negative prediction error, respectively (see Fig. 4 in **Supplementary Data Analysis** for trajectories of hyperparameters for a single subject):

$$\eta_{s,a}(t+1) = \begin{cases} \eta_{s,a}(t)+1 & \text{if } \delta_{s,a,t>0} \\ \eta_{s,a}(t) & \text{otherwise} \end{cases}$$

$$\beta_{s,a}(t+1) = \begin{cases} \beta_{s,a}(t)+1 & \text{if } \delta_{s,a,t<0} \\ \beta_{s,a}(t) & \text{otherwise} \end{cases}$$

The participant can then compare the means of each posterior distribution and adjust RTs so as to increase the probability of obtaining a reward prediction error. The mean of the beta distribution is simply $\mu = \eta/(\eta + \beta)$. Thus, this component of the exploitation model predicts that subjects adjust RTs according to $\rho[\mu_{\text{slow}}(s,t) - \mu_{\text{fast}}(s,t)]$, where ρ is a free parameter scaling the degree to which participants use these mean estimates in adapting their RTs.

In addition to the Go/NoGo learning and Bayesian integration mechanisms, model fits to data were also substantially improved by a mechanism in which participants adapted RTs toward that which had produced the single largest reward thus far ('going for gold'), regardless of the reward probability. This tendency was captured by free parameter υ and was not associated with any genotype (nor was it required for the core results of the paper to hold, but it may be useful for future studies of the neural and genetic mechanisms of this behavior). We modeled this by keeping track of the RT that yielded rewards that were at least 1 s.d. greater than all rewards observed thus far in the block and adapting all subsequent RTs toward this value. Further, participants' response on one trial may be heavily influenced by that of the previous trial, independent of value. Accordingly, we introduce a parameter λ to capture individual differences in this influence of previous responses.

Thus, the full RT model is as follows:

$$\hat{\text{RT}}(s,t) = K + \lambda \text{RT}(s,t-1) - \text{Go}(s,a,t) + \text{NoGo}(s,a,t) + \rho[\mu_{\text{slow}}(s,t) - \mu_{\text{fast}}(s,t)] + \upsilon[\text{RT}_{\text{best}} - \text{RT}_{\text{avg}}] + \text{Explore}(s,t)$$

The computations of the final Explore term is discussed next.

One of the central advantages of the Bayesian framework is that it provides an estimate not only of the 'best guess' (the mean, or expected value μ of the beta distribution) but also the uncertainty about that mean, quantified by the s.d. σ of that distribution. We attempted to predict RT swings from one trial to the next, hypothesizing that RT swings reflect exploration when participants are uncertain about whether they might obtain better outcomes. The s.d. of the beta distributions for each state (clock-face) can be computed analytically in each trial as a measure of uncertainty:

$$\sigma_{s,a}(t) = \sqrt{\left(\frac{\eta_{s,a}(t)\beta_{s,a}(t)}{(\eta_{s,a}(t)+\beta_{s,a}(t))^2(\eta_{s,a}(t)+\beta_{s,a}(t)+1)}\right)}$$

The model Explore term was applied on each trial as a function of the relative differences in uncertainty about the likelihood of reward prediction errors given fast and slow responses:

$$\text{Explore}\,(s,t) = \varepsilon[\sigma_{\delta|s,a=\text{Slow}} - \sigma_{\delta|s,a=\text{Fast}}]$$

In this way, exploratory-based RT swings are predicted to occur in the direction of greater uncertainty (thereby acting to reduce this uncertainty). Note that for trials immediately following an exploratory RT swing, as it stands this implementation would roughly double-count exploration because the λ parameter already reflects autocorrelation between the previous and current RT (where in this case the previous trial was an exploratory swing). To mitigate against this double counting, we set the Explore term to 0 in trials immediately following an exploratory RT swing (defined as a change in RT that was in the same direction predicted by the uncertainty Explore term). The results were not sensitive to this particular implementation, however. (For example, similar findings were found without resetting Explore to 0 but instead including a parameter into the RT estimate that reflects the effects of previous RT swings from trial $n-2$ to $n-1$ (in addition to λ, which accounts for the raw RT in trial $n-1$). This additional parameter was negative, such that a large RT swing in trial $n-1$ was predictive of a swing in the opposite direction in trial n. In this model, without Explore being re-set, all genetic findings remained significant, including the *COMT* gene-dose Explore effect; $P = 0.01$.)

A number of models of RT swings were compared in an effort to determine whether *COMT* effects were specific to uncertainty.

Sutton (1990) exploration bonus. In this model, exploration is increasingly encouraged for options that had not been explored for several trials. Specifically, exploration is predicted to increase with the square root of the number of trials since making that choice, scaled by free parameter ζ:

$$\hat{\text{RT}}'(s,t) = \begin{cases} \hat{\text{RT}}(s,t) + \varsigma\sqrt{(n)} & \text{if RT}(s,t-1) \ldots \text{RT}(s,t-n) < \text{RT}_{\text{avg}}(t-i) \\ \hat{\text{RT}}(s,t) - \varsigma\sqrt{(n)} & \text{otherwise} \end{cases}$$

"Lose-switch" model. In this model, RT swings are predicted to occur after negative prediction errors, such that participants switch to a slower response if the previous response was fast and vice versa. The degree of adaptation was scaled by free parameter κ.

$$\hat{\text{RT}}'(s,t) = \begin{cases} \hat{\text{RT}}(s,t) + \kappa & \text{if } \delta_{s,a,t-1<0};\ \text{RT}(s,t-1) < \text{RT}_{\text{avg}}(t-1) \\ \hat{\text{RT}}(s,t) - \kappa & \text{if } \delta_{s,a,t-1<0};\ \text{RT}(s,t-1) \geq \text{RT}_{\text{avg}}(t-1) \\ \hat{\text{RT}}(s,t) & \text{otherwise} \end{cases}$$

"Regression to the mean" model. Here responses are predicted to speed or slow as a function of whether the previous response was faster or slower than the local mean, regardless of the outcome. The degree of adaptation was scaled by free parameter ξ.

$$\hat{\text{RT}}'(s,t) = \begin{cases} \hat{\text{RT}}(s,t) + \xi & \text{if RT}(s,t-1) < \text{RT}_{\text{avg}}(t-1) \\ \hat{\text{RT}}(s,t) - \xi & \text{if RT}(s,t-1) \geq \text{RT}_{\text{avg}}(t-1) \end{cases}$$

where $\text{RT}'(s,t)$ is the new RT prediction including regression to the mean.

"Reverse momentum" model. This model attempts to capture periodic changes in RT whereby subjects reverse the direction of their responses if they had progressively sped up or slowed down over the last number of trials. The degree of RT adjustment was predicted to linearly increase with the number of preceding responses that had been progressively speeded or slowed, and scaled by a free parameter γ. Further, this RT reversal was predicted to occur only if the number of progressively speeded or slowed responses exceeded a minimum

threshold θ, also a free parameter (this parameter allows for variability in the period of RT swings and was required for the good fits described below).

$$\hat{RT}'(s,t) = \begin{cases} \hat{RT}(s,t) + \gamma n & \text{if } RT(s,t-1) < RT(s,t-2) < \ldots RT(s,t-n) \ldots; n > \theta \\ \hat{RT}(s,t) - \gamma n & \text{if } RT(s,t-1) > RT(s,t-2) > \ldots RT(s,t-n) \ldots; n > \theta \\ \hat{RT}(s,t) & \text{otherwise} \end{cases}$$

Model comparison results are presented in the **Supplementary Data Analysis**.

50. Ye, S., Dhillon, S., Ke, X., Collins, A.R. & Day, I.N. An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.* **29**, e88-1–e88-8 (2001).

Corrigendum: Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation

Michael J Frank, Bradley B Doll, Jen Oas-Terpstra & Francisco Moreno

Nat. Neurosci. **12, 1062–1068 (2009); published online 20 July 2009; corrected after print 9 September 2009**

In the version of this article initially published, the last sentence of the second paragraph in the right column on page 1065 read "that is, the following term was added to the RT prediction: $\rho[\sigma_{slow}(s,t) - \sigma_{fast}(s,t)]$, where ρ is a free parameter." A variable in the equation contained in this sentence was incorrect. The sentence should read "that is, the following term was added to the RT prediction: $\rho[\mu_{slow}(s,t) - \mu_{fast}(s,t)]$, where ρ is a free parameter." The error has been corrected in the HTML and PDF versions of the article.