Supplemental Information

The following information is included in the Supplemental file for this manuscript: 1) Additional detail regarding methods used in the computational model, 2) Behavioral results related to the role of antipsychotic medication in reinforcement learning, 3) Behavioral results related to frequency vs. magnitude bias, 4) Correlations between reinforcement learning parameters in the computational model and symptoms, and 5) Additional analyses indicating the specificity of a deficit in uncertainty-driven exploration in schizophrenia and its relationship with anhedonia.

Supplemental Methods and Materials

Computational Model

The model assumes that participants track the 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 controlling the degree to which values are updated on each trial, and δ is the reward prediction error signaled by midbrain dopamine neurons (1, 2), which is simply the reward outcome minus the prior expected value.

$$\delta(t) = Rew(t) - V(t)$$

Previous computational modeling in healthy participants identified multiple factors that govern trial-by-trial RTs, all of which capture non-overlapping variance in this task even when penalizing model fits for the inclusion of multiple parameters (3):

- An intercept K indicating participants' baseline motor response tendency independent of

other factors;

- A parameter ρ which predicts that individuals adjust RTs in the direction of greater probability of obtaining a positive outcome based on the observed reward statistics. Bayes' rule is applied to track the probability of obtaining a positive prediction error, separately for fast (RT < median) or slow (RT > median) responses. (Only two response categories are needed because the reward functions are monotonic: any asymmetry in rewards for fast vs. slow responses can be capitalized by simply adjusting RTs in the direction of greater reward probability.) These probabilities are updated as a function of each outcome:

$$P(\theta|\delta_1 ... \delta_n) \propto P(\delta_1 ... \delta_n|\theta)P(\theta),$$

where θ reflects the parameters governing the belief distribution about likelihood of reward prediction errors for each response, and δ1...δn are the prediction errors observed thus far (on trials 1 to n).

The belief distribution is initialized to be uniform at the beginning of each block of trials. The posterior distribution is calculated after each response and outcome. At each trial, response times are then predicted to vary as a function of the difference between the reward probability estimates (means μ of the beta belief distributions) for fast and slow responses. The fast and slow RT split is an assumption that is justified for several reasons. First, because reward functions are monotonic in this task, all one has to do is keep track of reward statistics for fast and slow responses and adjust RTs in the direction of that with the highest expected value. Keeping track of just two response-reward associations is perhaps more plausible than separately maintaining reward statistics for multiple RTs, particularly when the task requires comparing the expected values (and uncertainties) of these different responses. Finally, modeling the task in this way (with both the mean value and uncertainties of each response category) allows us to provide a reasonable fit to the data that is improved relative to not incorporating either of these factors. Nevertheless, the free response time allows us to assess directional changes in RT consistent with that

expected by incremental adjustments or exploration, even if the change in RT does not always correspond to a categorical shift from fast to slow or vice-versa.

- An exploration parameter ϵ predicts trial-by-trial RT swings to occur when one is relatively more uncertain about the reward probabilities for fast or slow responses. The standard deviations σ of the above beta distributions are computed on each trial as estimates of outcome uncertainty for each response category. In particular, the Explore term of the model is computed as follows:

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

where ϵ is a free parameter that scales the degree of exploration in proportion to relative uncertainty and $\sigma_{\delta|s,a=Slow}$, $\sigma_{\delta|s,a=Fast}$ are the standard deviations quantifying uncertainty about positive outcomes given slow and fast responses, respectively. These uncertainty estimates generally decrease with more evidence (i.e., one becomes more confident about the reward probabilities for fast responses after making several fast responses), albeit at a slower rate for more variable outcomes. 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;

- Two learning rate parameters, α_G and α_N, estimating the degree to which individuals speed and slow RTs as a function of accumulated positive and negative prediction errors, respectively. In contrast to the Bayesian statistical learning process (captured by parameter ρ), these parameters capture a more implicit bias to produce speeded responses following multiple positive prediction errors, and slowed responses following multiple negative prediction errors. These are thought to capture basal ganglia dopaminergic mechanisms modulating corticostriatal synaptic plasticity in the Go and NoGo pathways, promoting approach and avoidance responses respectively, and consistent with available genetic and pharmacological data in this task (3, 4);

- A response recency parameter λ scaling the impact of the previous response's RT on the current choice, independent of any change in value (see also similar parameters in other reinforcement tasks (5));
- A "going for gold" parameter ν predicting that participants will adjust RTs toward that which has produced the single largest reward experienced thus far, and which is at least one standard deviation greater than the other rewards, regardless of the probability of occurrence. As noted above, each of these parameters captures variance in this task beyond that of the others (3). The complete RT model for each clock face state s and trial t is thus as follows:

$$\begin{aligned}\hat{RT}(s,t) = \quad & K + \lambda RT(s,t-1) - Go(s,t) + NoGo(s,t) \\ & + \rho[\mu_{slow}(s,t) - \mu_{fast}(s,t)] + \nu[RT_{best} - RT_{avg}] \\ & + Explore(s,t),\end{aligned}$$

where $Go(s,t)$ and $NoGo(s,t)$ terms are the cumulative sums of positive and negative prediction errors, scaled by αG and αN, respectively. Complete model details are given in (3).

In all models, we used the Simplex method (6) with multiple starting points to derive best fitting parameters for each individual participant that minimized the sum of squared error between predicted and actual RTs across all trials, $\Sigma_t(\hat{RT} - RT)^2$. A single set of parameters was derived for each subject providing the best fit across all task conditions.

Supplemental Results

Descriptive Statistics for Primary Behavioral Task Conditions

Table S1. Means and SDs for CEV, DEV, and IEV conditions in patients (SZ) and controls (CN).

	CN (n = 39)	**SZ (n = 51)**
CEV	2277 (588)	2199 (644)
DEV	1989 (557)	1992 (623)
IEV	2518 (594)	2454 (628)

Values reflect mean RT scores (SD) averaged across all trials for each condition.

We also examined the potential for Group differences in Go and No Go learning in DEV and IEV conditions after controlling for baseline differences in CEV performance. Using DEV-CEV and IEV-CEV difference scores, a 2 Group (SZ vs. CN) X 2 Condition (DEV-CEV vs. IEV-CEV) repeated measures ANOVA indicated a nonsignificant interaction, when calculated using the second half of trials per block ($F = 1.46$, $p = 0.23$).

Frequency vs. Magnitude Bias

To determine whether patients and controls differed with regard to showing a greater bias toward learning about reward probability (frequency) or magnitude, we examined group differences in behavioral performance in the CEVR condition. Specifically, we calculated a probability vs. magnitude bias score by taking the difference in RT between the CEVR and CEV conditions (P-M bias = CEVR-CEV). In both groups, RTs were longer in the CEVR than CEV condition, CN = 303 ms (602 ms); SZ = 300 ms, (905 ms); however, a one-way ANOVA indicated that there were no significant differences between SZ and CN on the P-M bias estimate, $F (1, 88) < 0.01$, $p = 0.987$ ($\eta 2 = 0.00$). This indicates that both groups have a bias toward higher reward frequency.

We also examined frequency vs. magnitude learning in relation to negative symptoms. Figure S1 presents frequency-magnitude bias data for HI-NEG, LOW-NEG, and CN subjects, and indicates that both the LOW-NEG and CN subjects show the general pattern of having a bias toward learning more about reward probability than magnitude (as indicated by the positive difference score), while the HI-NEG patients fail to show a bias toward either probability or magnitude. One-way ANOVA supported this interpretation, indicating a significant difference among the 3 groups on the probability vs. magnitude bias measure (P-M bias = CEVR-CEV), $F(2, 85) = 3.27$, $p = 0.04$ ($\eta 2 = 0.07$). Post hoc Scheffe contrasts indicated significant differences between the HI-NEG and LOW-NEG patients ($p = 0.045$), but not between the HI-NEG and CN ($p = 0.234$) or LOW-NEG and CN ($p = 0.558$). Thus, while the LOW-NEG and CN groups both show a pattern toward learning more about probability than magnitude, the P-M contrast only reached statistical significance between HI-NEG and LOW-NEG groups. This is further displayed by the correlation between SANS total negative symptoms and the P-M contrast, which was at a trend level of significance ($r = -0.26$, $p = 0.08$). Overall, these results suggest that patients were less sensitive to reward magnitudes than probabilities, but counter-intuitively, those with the most severe negative symptoms showed relatively greater sensitivity to reward magnitudes. To determine whether the computational modeling analyses might shed light on this issue, we conducted an exploratory analysis in which we correlated other model parameters with SANS total scores (despite the fact that, unlike αG, these other parameters did not differ overall between patients and controls). In this analysis, we found only one parameter to correlate with SANS Total: the “going for gold” parameter. This positive correlation indicates that patients with the greatest degree of negative symptoms were more likely to adjust their response time toward that which had previously yielded an exceptionally large reward magnitude, regardless of its probability of occurrence. This finding may indicate that those patients with the most severe negative symptoms are simply insensitive to all reward outcomes except those having the very largest values – i.e., they may have a higher threshold for considering an outcome relevant.

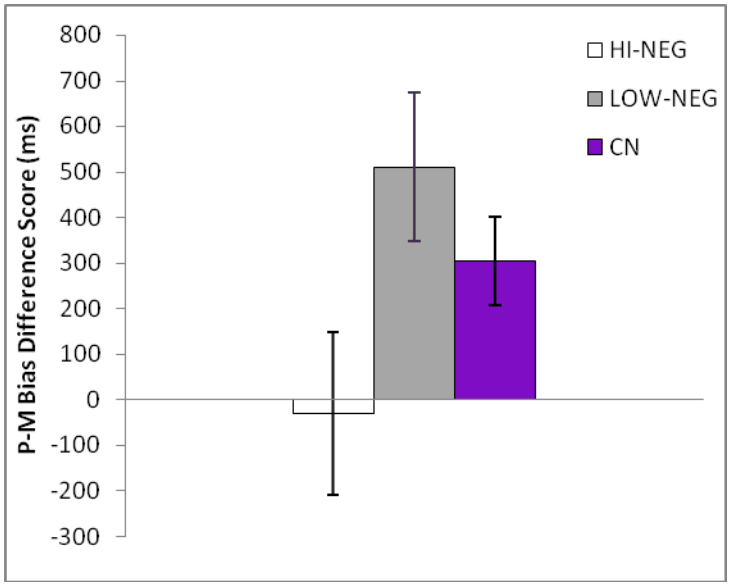

Figure S1. Frequency vs. magnitude bias in HI-NEG, LOW-NEG, and CN subjects. Mean RT difference score between CEVR and CEV conditions, reflecting potential biases in learning more about reward frequency or reward magnitude. Extreme positive difference scores reflect a bias toward learning more about reward frequency then magnitude, values near 0 reflect the lack of a bias toward frequency or magnitude, extreme negative difference scores reflect a bias to learn more about magnitude then frequency.

The Role of Antipsychotic Medications in Reinforcement Learning

To further examine the effects of antipsychotic medication, patients were divided into medication groups based upon low potency and high potency D2 blockade. Low potency drugs included clozapine, quetiapine, and olanzapine. High potency drugs included aripiprazole, haloperidol, risperidone, fluphenazine, and ziprasidone. Patients on multiple medications were excluded from these analyses (n = 10). Repeated measures ANOVA indicated a nonsignificant Medication Group (high vs. low) X Condition (DEV, IEV) interaction ($F = 0.15$, $p = 0.70$). Individual one-way ANOVAs also indicated that the groups failed to differ in the DEV ($F = 0.87$, $p = 0.36$), IEV ($F = 2.10$, $p = 0.16$), or P-M contrast ($F = 2.33$, $p = 0.14$) conditions. Thus, when patients are divided into antipsychotic medication groups, there appears to be little difference in reinforcement learning (see Table S2).

Table S2. Mean (SD) RT (ms) Performance on Reinforcement Learning Conditions in Patients in Low Potency and High Potency D2 Blocking Antipsychotics

	Low Potency D2 Blockers (n = 32)	**High Potency D2 Blockers (n = 9)**
DEV Change Score (Go Learning)	-130 (441)	29 (495)
IEV Change Score (No Go Learning)	13 (487)	274 (431)
P-M Bias (Probability vs. Magnitude)	385 (833)	-148 (1225)

Correlations Between Reinforcement Learning Model Parameters and Symptoms

Spearman correlations were calculated between modeling parameters and the SANS total score, BPRS positive, BPRS disorganized, and BPRS total symptom scores to determine the specificity of symptom associations. Only the aforementioned correlation between "going for the gold" and SANS total was significant. The likely interpretation for why the exploration parameter correlated with SANS anhedonia, but not the BPRS negative syndrome score is because the BPRS does not include items for anhedonia or avolition. The lack of correlations with other symptom domains reflect the specificity of the association with anhedonia (which as rated by the SANS is quite distinct from related concepts like depression).

Model-Fits

Model fits evaluated with the Akaike Information Criterion (AIC) did not differ across groups. There was a nonsignificant trend for fits to be improved in patients (mean AIC = 2955) relative to controls (mean AIC 3035, $p = 0.08$), possibly due to the reduced exploratory RT swings in patients which are difficult for the model to fit and cause penalties in squared error (which are partially but imperfectly mitigated by the model explore term). Figure S2 depicts example single subject data, showing the correspondence of the model Explore term with trial-by-trial RT swings (similar to 3).

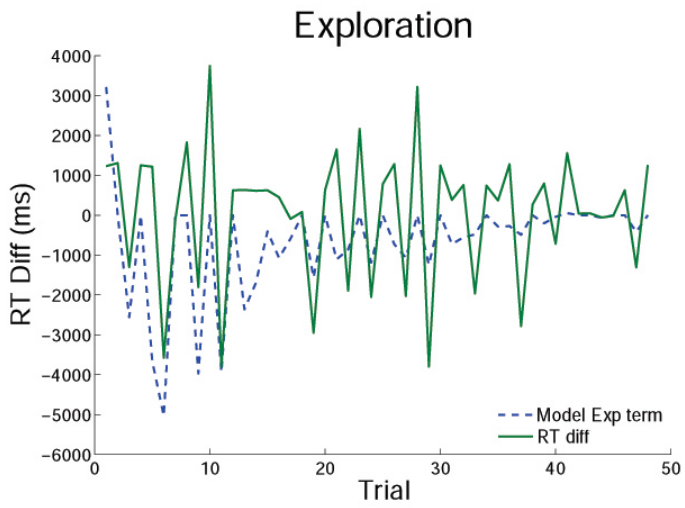

Figure S2. Example single subject data showing correspondence between Explore parameter and trial-by-trial RT Swings

Figure S3 displays probability distributions of experienced reward statistics representing mean and variance of expected outcomes. Plots A and B depict beta probability density distributions representing the belief about the likelihood of achieving a better than expected outcome for fast and slow responses, respectively. The x-axis is 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. Differences between the μfast and μslow were used to adjust RTs to exploit rewarding responses. Means evolve to be higher for fast responses in DEV and for slow responses in IEV. The standard deviation σ, which decreases as a function of experience, was taken as an index of uncertainty. Exploration was predicted to modulate RT in direction of greater uncertainty in a given trial about whether outcomes might be better than the status quo.

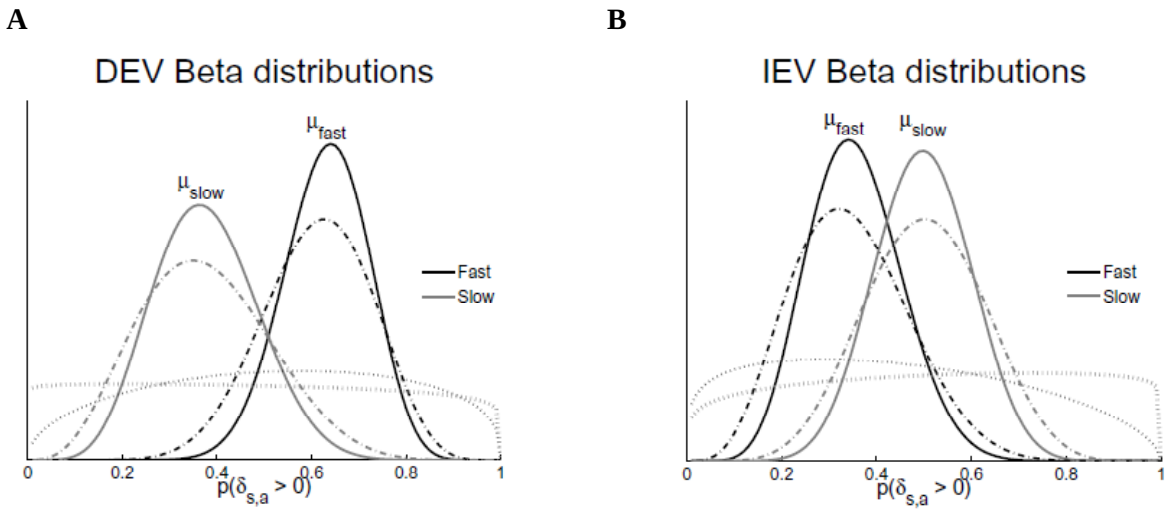

Figure S3. Probability distributions of experienced reward statistics representing mean and variance of expected outcomes

Uncertainty-Based Exploration

We also examined the relationship between uncertainty-driven exploration and negative symptoms. However, nearly one half of the patients (23/51, compared with 8/39 controls) were best fit by an exploration parameter of 0. (This proportional group difference in number of subjects with non-zero ϱ was also significant; two-tailed Fisher's exact test p = 0.02). This was caused by the fact that we had constrained the exploration parameter to be positive in the first run, i.e., it would only predict a change in RT in proportion to greater uncertainty, and if subjects did not employ this strategy their exploration term would be zero. Because many of the patients had zero exploration terms (consistent with the significant reduction in exploration in patients) it was not possible to correlate this term with symptoms within the patient group, due to relative lack of variability. This lack of variability precluded us from properly investigating individual differences as a function of avolition/anhedonia in the basic model. To address this issue, we re-ran the model fits without enforcing 0 as the lower bound for exploration, but instead allowing it to reach a negative value. Negative values would be expected if participants are more likely to avoid actions with uncertain outcomes rather than to explore them, and to instead continue to make those

responses for which reward statistics are more certain. Thus, individuals with more negative exploration terms would be that much more unlikely to produce exploratory responses in proportion to uncertainty. We verified that this analysis produced interpretable results in our original genetics study, in which explore values were bounded at zero (3). We confirmed that the significant gene-dose effect of the COMT gene continued to hold in this analysis (with met/met showing the highest exploration and val/val the lowest; $p < 0.05$). Only here, participants in the val/val genotype exhibited, on average, negative explore values – confirming that they were particularly unlikely to explore uncertain actions.

When we refit model parameters across all subjects, allowing for negative exploration, the same effects, and their significance emerged (explore $p < 0.02$; α G $p = 0.07$; logistic explore $p = 0.02$, α G $p = 0.028$), with no difference in any other model parameters. Allowing the exploration term to reach negative values thus revealed individual differences such that those with more negative values are that much less likely to explore, and that much more likely to stick with certain choices. That the main finding between patients and controls replicates when this term is allowed to be negative in both groups makes us more confident that the overall patient reduction in uncertainty-driven exploration is reliable and not dependent on this implementational detail.

Additionally, given the unique association with anhedonia, we further investigated whether the impact of anhedonia on exploration was specific to uncertainty. First, we computed whether increasing anhedonia was simply associated with reduced overall RT variability, which might lead to reduced ϵ parameter estimates without necessarily implying that patients are less sensitive to uncertainty about possible benefits of exploratory actions. The data are not consistent with this interpretation, however, as RT variability (as estimated by the coefficient of variation, which is the standard deviation normalized by mean RT (7, 8) did not correlate with anhedonia ($p > 0.5$). We further computed a measure of trial-to-trial variance as a measure of whether overall RT swings were smaller as a function of anhedonia. To this end, we computed consecutive RT variance:

$$\sqrt{(\sum (RT(i) - RT(i+1))^2/(n-1))},$$

where *i* is trial number and *n* is the total number of trials (7). This measure of consecutive RT variance also showed no correlation with anhedonia (r = 0.05, ns). Thus anhedonia was not associated with an overall reduction in RT swings; rather, RT swings were less likely to vary in the direction of relative uncertainty. RT variability did not differ between patients and controls; however, patients had significantly lower consecutive variance scores, suggesting that patients have smaller RT swings than controls ($p < 0.001$).

In addition, to determine whether any effects of schizophrenia and avolition/anhedonia on exploration were specific to uncertainty, we also included other alternative models found in prior work to account for other variance in RT swings, including a regression to the mean parameter and a lose-switch parameter (3):

- Regression to the mean parameter ξ. RTs are predicted to speed up after slow responses and slow down after fast responses, regardless of outcome:

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

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

- Lose-switch parameter κ. 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.

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

If schizophrenia effects are specific to uncertainty-driven exploration, they would then be observed on parameter ϵ but not ξ or κ. Two follow-up simulations confirmed a selective effect of

uncertainty. In these simulations, we included additional parameters that capture trial-by-trial RT swings but which are insensitive to uncertainty (regression to the mean parameter ξ and lose-switch parameter κ; see also 3)). In these models, the exploration parameter was still significantly lower in patients relative to controls ($p = 0.01$ and $p = 0.03$, respectively), with no effect on any other parameter, including ξ or κ. That these effects are specific to uncertainty is supported by an analysis in which we computed relative parameter estimates for exploration compared to κ or ξ (with each parameter converted to standardized z-scores so that that they are on the same scale). Relative exploration scores indicate the extent to which RT swings are affected by uncertainty as compared to these other factors. In this analysis, relative exploration values were again significantly reduced in patients compared to controls (both p's = 0.01). Moreover, anhedonia continued to significantly correlate with relative reductions in exploration (p's < 0.01). Further, with either of these parameters included into the model (capturing additional variance), the αG parameter difference between patients and controls reached significance or near-significance ($p = 0.057$ and $p = 0.04$), but still did not correlate with anhedonia. Additionally, κ and ξ were not significantly correlated with anhedonia (p's > 0.39), suggesting that the correlation with anhedonia is unique to uncertainty-driven exploration.

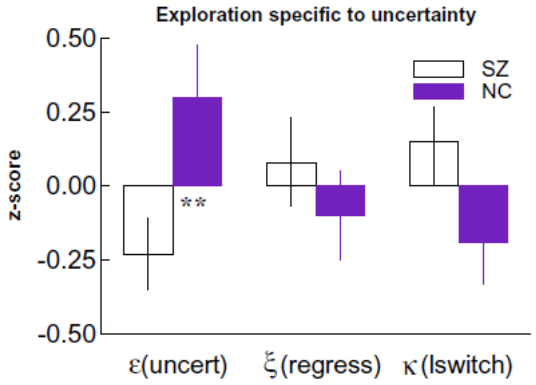

Figure S4. Differences in exploration parameters between schizophrenia patients and controls. The figure depicts standardized z-scores for the uncertainty-driven exploration parameter ϱ and alternative models of trial-to-trial dynamics (regression to the mean, and lose-switch) demonstrate that SZ deficiencies are specific to uncertainty. Error bars reflect standard error.

Supplemental References

1. Bayer HM, Glimcher PW (2005): Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron.* 47(1):129-41.
2. Montague PR, Dayan P, Sejnowski TJ (1996): A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci.* 16(5):1936-47.
3. 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(8):1062-8.
4. Moustafa AA, Cohen MX, Sherman SJ, Frank MJ (2008): A role for dopamine in temporal decision making and reward maximization in parkinsonism. *J Neurosci.* 28(47):12294-304.
5. Lau B, Glimcher PW (2007): Action and outcome encoding in the primate caudate nucleus. *J Neurosci.* 27(52):14502-14.
6. Nelder JA, Mead R (1965): A simplex method for function minimization. *Computer Journal*. 7: 308–313.
7. Klein C, Wendling K, Huettner P, Ruder H, Peper M (2006): Intra-subject variability in attention-deficit hyperactivity disorder. *Biol Psychiatry*. 60(10):1088-97.
8. Frank MJ, Santamaria A, O'Reilly RC, Willcutt E (2007): Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology.* 32:1583–1599.