Supplemental Material

Baseline dopamine predicts drug effects on reversal learning

Supplemental Methods

Computational reinforcement learning model

In the model, the value *V* of the highlighted stimulus *i* was updated according to the following rule after each trial:

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

and the value of the alternative stimulus *j* was also updated in the opposite direction:

$$V_j(t+1) = V_j(t) - u\alpha\delta$$

where δ is the predicton error, $r(t) - V_i(t)$, α is a learning rate and *u* is an update parameter that determines the extent to which participants update the value of the alternative stimulus by inference (see below). The learning rate applied depends on the outcome, and is α_R for unexpected rewards ($\delta > 0$) and α_P for unexpected punishments ($\delta < 0$) (Frank et al., 2007). Choice behaviour was then modelled using a standard "softmax'' logistic function (Sutton and Barto, 1998), with inverse gain parameter β, such that the probability of predicting reward was computed as a function of the relative difference in value for the highlighted stimulus *i* compared to the alternative stimulus *j*:

$$P_R(t) = 1/(1 + \exp(-(V_i(t) - V_j(t))/\beta)$$

and the probability of predicting punishment $P_P(t)$ is just 1 – $P_R(t)$. We then maximized the log likelihood for each participant's behavioural data by finding the best fitting parameters α_R, α_P and *u* that corresponded to their trial-by-trial sequence of choices across all blocks within each behavioural session (see for example Daw et al., 2006; Frank et al., 2007).

Note that we set the value of β to a constant across all subjects, because this was a parameter unrelated to the specific hypotheses, and to restrict the total number of free parameters to a minimum. Nevertheless we found the best fitting β across the group, which was 0.3. Our simulations confirmed that holding β constant across the group and allowing the update parameter *u* to vary as a free parameter provided a substantially better fit than the other way around. This in an effort to minimize free parameters we fixed β and allowed others to vary.

We applied a rule for updating the value of the alternative (non-highlighted) stimulus in the opposite direction (Matsumoto and Hikosaka, 2007). This double update model captures the knowledge of higher order task structure in our deterministic reversal task, and is similar in essence to the state-based model of Hampton et al (2006) which captured this same knowledge of reversal and provided a substantially better fit to behaviour than a simple, single update RL model. Given that the task here is deterministic, there was no need to apply Bayesian analysis for computing the posterior probability that a given unexpected outcome is accompanied by a state reversal.

We first confirmed the model fit to the data, quantified by pseudo-R^2, which compares log likelihood of data under the model compared to a model that predicts random choice (P_R = 0.5 for all trials) (Camerer and Ho, 1998) was adequate.

Indeed, the mean pseudo-R^2 value was 0.69, which is quite high for model fits to individual trial-by-trial choices (compare with Daw et al., 2006; Frank et al., 2007), and substantially higher than a standard RL model that does not include reversal structure (Hampton et al., 2006). Perhaps this fit is not surprising given the deterministic nature of the task and the lack of exploration; nevertheless, the fit is substantially higher than that derived from the best fitting single update models (mean pseudo-R^2 = 0.45), easily justifying the use of the extra *u* parameter even when penalizing the model for having this parameter using Aikake's Information Criterion (AIC; Burnham and Anderson, 2002). Further, the mean update parameter *u* was 0.85, suggesting that participants were likely to update the value of the alternative stimulus after unexpected outcomes nearly as much as they updated the value of the highlighted stimulus in the trial.

As for accuracy, relative (difference) scores were calculated by subtracting punishment learning rates from reward learning rates.

Supplemental Results A

The behavioural data reported in the main text indicate that individual differences in reward- versus punishment-based reversal, and their sensitivity to D2 receptor stimulation are highly dependent on baseline dopamine synthesis capacity in the human striatum. This dependency, however, might reflect modulation of one of two fundamental processes. First, the effects might reflect modulation of the need to overcome and/or inhibit well-established predictions, given that contingency reversals and thus the critical switch trials occurred only after attainment of a learning criterion. Second, they might reflect more general associative learning mechanisms that happen to have surfaced most readily on switch trials due to the disproportionate unexpectedness of the outcome preceding those switch trials. To

disentangle these alternative hypotheses, we applied computational reinforcement learning algorithms that allowed the generation of learning-rate parameters (separately for reward and punishment) (see above). These learning-rates reflect updating not only on switch trials but also on other trials when learning criteria have not yet been attained and contingencies have presumably been less well established. We then investigated whether individual differences in these learning-rate parameters derived from the computational model could be accounted for by baseline dopamine synthesis capacity in the striatum.

The results from the model-based analyses parallel the pattern of results from the trial-based analyses. However, the model-based analyses also revealed additional effects (see below).

We ran a repeated measures ANOVA on the best fitting learning rate parameters for each subject under placebo with valence as a within-subject factor and striatal dopamine synthesis capacity and acquisition delay as covariates. This analysis revealed a significant valence by synthesis capacity interaction ($F_{1,8} = 8.4$, $P = 0.02$), which was due to a significant positive correlation between striatal dopamine synthesis and relative learning rates (α_R values minus α_P values) ($r_8 = 0.71$, $P = 0.02$). Thus, subjects with greater dopamine synthesis showed greater learning from unexpected reward relative to learning from unexpected punishment. We then examined whether baseline synthesis rates also predicted the effects of bromocriptine on the learning rate parameters derived from the computational model. Consistent with the behavioural data (which were evaluated only on switch trials), there was a significant three-way interaction between drug, valence and synthesis capacity ($F_{1,7} = 19.3$, $P = 0.003$), which was due to a negative correlation between dopamine synthesis capacity and drug-induced changes in relative learning

rates (r_7 = -0.86, P = 0.003). Critically, a significant relationship was obtained between dopamine synthesis and the drug effect on reward learning rate (r_8 = -0.71, P = .02) (**Supplemental Figure a**), as well as between dopamine synthesis and the drug effect on punishment learning rate (r_7 = 0.78, P = 0.013) (**Supplemental Figure b**).

It might be noted that there was relatively little variability and a possible ceiling effect in the absolute learning rates (**Supplementary Table 3**), derived from the model. However, this issue was resolved when scores were expressed in terms of drug effects, the effects of interest in the current study (**Supplemental Figure**).

The finding that dopamine synthesis capacity correlated significantly with the drug effect on reward learning rate, but not with that on reward-based reversal accuracy highlights the added value of the model-based analyses, which made use of information that is not directly observable in the behavioural data. Furthermore, the finding that correlations with learning rates were significant, whereas those with performance were not, indicates that the effects are unlikely to reflect switch-specific processes. Indeed, incorporation of model-derived information from non-switch trials increased rather than decreased the sensitivity of the dependent measure, suggesting that the effects reflect a more general associative learning mechanism.

Supplemental Results B

Previous studies have revealed that the effects of bromocriptine can be predicted from baseline working memory capacity, as measured with the listening span test (Kimberg et al., 1997; Gibbs and D'Esposito, 2005; Frank and O'Reilly, 2006; Cools et al., 2008). Consistent with the present results, we recently reported that the listening span correlates with striatal dopamine synthesis capacity (Cools et al.,

2008). These findings led us to investigate whether the drug effects on reversal learning, which correlated with dopamine synthesis capacity, also depended on baseline working memory capacity, as measured with the listening version (Salthouse and Babcock, 1991) of a reading span task modeled after Daneman & Carpenter (1980). Consistent with our prediction, span predicted drug effects on learning, although span-dependency was restricted to drug-induced changes in punishment-based reversal learning (accuracy: $r_7 = 0.78$, $P = 0.013$; learning rate: $r_7 = 0.72$, $P = .03$) and did not extend to drug-induced changes in reward-based reversal learning (accuracy: $r_8 = 0.18$, ns; learning rate: $r_8 = -0.1$; ns). Interestingly, the partial correlation between span and drug-induced changes in punishment-based reversal learning, while controlling for dopamine synthesis capacity, was no longer significant ($r_7 = 0.57$, $P = 0.1$). This observation is consistent with the hypothesis that the effects of span reflect effects of baseline dopamine function. Thus working memory span is a valid predictor of the effects of dopaminergic drugs on reversal learning, consistent with neurochemical data linking span to baseline differences in striatal DA synthesis (Cools et al., 2008) as well as previous behavioural data (Frank and O'Reilly, 2006).

Supplemental Results C

The delay between the acquisition of the PET data and that of the behavioural data was considerable. However, we argue that this delay does not confound our results. Critically we have explicitly addressed in a quantitative manner the possibility that the effect reflects noise. To this end, we have added variability to each subject's Ki values by drawing from a random distribution with the same statistics published on inter-individual variability of PET measures (so that each subject's value remained within 18% of their original). We repeated this process 100 times. Despite this added variability, which poses an additional source of noise beyond that afforded by the

existing delay between measurements the correlation remained statistically significant at $P < 0.05$ in 80 out of 100 cases.

Supplemental Figure Legend

Baseline-dependency of drug effects on model-derived reward and punishment learning rates and their sensitivity to D2 receptor stimulation.

(A) Significant negative correlation between striatal dopamine synthesis capacity and the effect of bromocriptine on reward learning rate (bromocriptine minus placebo). (B) Significant positive correlation between striatal dopamine synthesis capacity and the effect of bromocriptine on punishment learning rate (bromocriptine minus placebo). For statistics, see text.

Supplemental Table 1

Number of completed stages within each condition

	Placebo	Bromocriptine
Unexpected reward	25.2 (1.0)	23.9 (1.4)
Unexpected punishment	26.4 (0.6)	27.4 (0.9)

Values (standard errors of the mean) represent mean number of stages completed as a function of condition and drug session. Values ranged from 18 to 31. ANOVA with drug and condition as within-subject factors and with dopamine synthesis capacity and acquisition delay as covariates revealed no effects of drug, condition, synthesis capacity, nor any interactions.

Supplemental Table 2 Mean proportion of correct responses after unexpected reward and after unexpected punishment

	Reward	**Punishment**
Placebo	0.91 (0.03)	0.95 (0.03)
Bromocriptine	0.92 (0.03)	0.93 (0.02)

Values represent means (standard errors of the mean)

Supplemental Table 3 Individual learning rates

Subject	α reward placebo	α punishment placebo	α reward bromocriptine	α punishment placebo
1	1	.95	1	1
2	1	1	1	na
3	.8	.9	1	.85
4	1	1	1	1
5	.9	.95	1	1
6	1	.95	.9	1
7	.85	1	.85	.9
8	1	1	1	1
9	.9	1	.85	1
10	1	.9	.9	1
11	.95	1	1	1

Supplemental References

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: A Practical-Theoretic Approach, 2nd ed. Springer-Verlag: Springer-Verlag.

Camerer C, Ho T (1998) Experience-Weighted Attraction Learning in Coordination Games: Probability Rules, Heterogeneity, and Time-Variation. J Math Psychol 42:305-326.

Cools R, Gibbs S, Miyakawa A, Jagust W, D'Esposito M (2008) Working memory capacity predicts dopamine synthesis capacity in the human striatum. J Neurosci 28:1208-1212.

Daneman M, Carpenter P (1980) Individual differences in working memory and reading. J Verbal Learning Verbal Behav 19:450-466.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876-879.

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

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

Gibbs SE, D'Esposito M (2005) Individual capacity differences predict working memory performance and prefrontal activity following dopamine receptor stimulation. Cogn Affect Behav Neurosci 5:212-221.

Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26:8360-8367.

Kimberg DY, D'Esposito M, Farah MJ (1997) Effects of bromocriptine on human subjects depend on working memory capacity. Neuroreport 8:3581-3585.

Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 447:1111-1115.

Salthouse T, Babcock R (1991) Decomposing Adult Age Differences in Working Memory. Dev Psychol 27.

Sutton R, Barto A (1998) Reinforcement learning. Cambridge, MA: MIT Press.

Vingerhoets F, Snow B, Schulzer M, Morrison S, Ruth T, Holden J, Cooper S, Calne D (1994a) Reproducibility of fluorine-18-6-fluorodopa positron emission tomography in normal human subjects. J Nucl Med 35:18-24.

Vingerhoets FJ, Snow BJ, Lee CS, Schulzer M, Mak E, Calne DB (1994b) Longitudinal fluorodopa positron emission tomographic studies of the evolution of idiopathic parkinsonism. Ann Neurol 36:759-764.