EJN fens European Journal of Neuroscience

European Journal of Neuroscience, Vol. 35, pp. 1024–1035, 2012 doi:10.1111/j.1460-9568.2011.07980.x

How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis

Anne G. E. Collins and Michael J. Frank
Department of Cognitive, Linguistic and Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence, RI, USA

Keywords: basal ganglia, capacity, human, prefrontal cortex, striatum

Abstract

Instrumental learning involves corticostriatal circuitry and the dopaminergic system. This system is typically modeled in the reinforcement learning (RL) framework by incrementally accumulating reward values of states and actions. However, human learning also implicates prefrontal cortical mechanisms involved in higher level cognitive functions. The interaction of these systems remains poorly understood, and models of human behavior often ignore working memory (WM) and therefore incorrectly assign behavioral variance to the RL system. Here we designed a task that highlights the profound entanglement of these two processes, even in simple learning problems. By systematically varying the size of the learning problem and delay between stimulus repetitions, we separately extracted WM-specific effects of load and delay on learning. We propose a new computational model that accounts for the dynamic integration of RL and WM processes observed in subjects' behavior. Incorporating capacity-limited WM into the model allowed us to capture behavioral variance that could not be captured in a pure RL framework even if we (implausibly) allowed separate RL systems for each set size. The WM component also allowed for a more reasonable estimation of a single RL process. Finally, we report effects of two genetic polymorphisms having relative specificity for prefrontal and basal ganglia functions. Whereas the COMT gene coding for catechol-O-methyl transferase selectively influenced model estimates of WM capacity, the GPR6 gene coding for G-protein-coupled receptor 6 influenced the RL learning rate. Thus, this study allowed us to specify distinct influences of the high-level and low-level cognitive functions on instrumental learning, beyond the possibilities offered by simple RL models.

Introduction

Over the past decade, the computational framework of reinforcement learning (RL) (Sutton & Barto, 1998) has enjoyed widespread use in the study of instrumental learning. This increasing popularity is rooted in two main distinct factors. The first is that RL is a very simple computational framework that can account for a variety of behavioral effects in learning experiments across species. The second is that this same framework has also provided a principled basis for understanding the properties of a wide range of neurobiological data. This framework has thus served as a highly successful methodological and conceptual tool for linking brain function to behavior.

More precisely, RL algorithms are thought to be implemented by the corticostriatal circuitry and its modulation by dopamine. In particular, phasic changes in the firing of dopaminergic cells (Montague *et al.*, 1996; Schultz, 1997; Bayer & Glimcher, 2005, etc.) have been shown to encode a signal corresponding to the crucial RL quantity of the reward prediction error, defined as the difference between the expected and observed values at a given point in time. Activity and plasticity in striatal neurons, a major target of dopaminergic efferents, are dynamically sensitive to these dopaminergic prediction error signals, which enable the striatum to represent RL values (O'Doherty *et al.*, 2004; Frank, 2005; Daw & Doya, 2006). Behavioral predictions derived from this mechanistic corticobasal ganglia framework have been widely confirmed in studies using a variety of methods including patients, pharmacology, gene studies, etc. (Frank & Fossella, 2010). Further, neural correlates of RL values and prediction errors have consistently implicated the striatum in monkey and rat electrophysiology and human functional imaging (O'Doherty *et al.*, 2004; Samejima *et al.*, 2005; Pessiglione *et al.*, 2006; Lau & Glimcher, 2007; Jocham *et al.*, 2011). Thus, although this remains an active research area and aficionados continue to debate some of the details, there is widespread general agreement that the basal ganglia (BG) and dopamine are critically involved in the implementation of RL.

It should be noted that human instrumental learning experiments do not rely exclusively on incremental learning from prediction errors, but probably involve a wider array of higher level function, including executive functions and working memory (WM). This can be seen in the wide recruitment of brain areas involved in such paradigms that strongly implicate the prefrontal cortex, especially during early learning. Thus, attempting to quantitatively capture instrumental learning will necessitate that the RL model accounts for the influences of these higher order cognitive mechanisms. It is particularly crucial to

Correspondences: Anne G. E. Collins and Michael J. Frank, as above.
E-mails: michael_frank@brown.edu and anne_collins@brown.edu

Received 26 October 2011, revised 28 November 2011, accepted 1 December 2011

be aware of this issue when RL models are applied for the model-based analysis of neuroscientific data, including electrophysiology (Corrado & Doya, 2007), functional magnetic resonance imaging (O'Doherty *et al.*, 2004, 2007; Seymour *et al.*, 2004; Tanaka *et al.*, 2004, 2006; Daw & Doya, 2006; Pessiglione *et al.*, 2006; Schönberg *et al.*, 2007; Brovelli *et al.*, 2008; Kahnt *et al.*, 2009), electroencephalography (Cavanagh *et al.*, 2010) and genetic data (Frank *et al.*, 2007, 2009a; Doll *et al.*, 2011).

Consider, for example, the RL model learning rate, which encodes the degree to which the prediction error leads to an adjustment of expected action values. Thus, using a simple RL framework, one could estimate the learning rate based on observed behavioral sequences of choices, and expect this to represent the efficacy of the dopamine action on the striatum [as one example, Kahnt *et al.* (2009) links the learning rate to midbrain–striatum connectivity]. However, because learning rates are fit to the behavioral organism as a whole, to the extent that behavioral adjustments during learning involve WM and hypothesis testing, these factors will greatly affect the estimated learning rate (Frank *et al.*, 2007). Indeed, associations between individual differences in learning rates and genetic factors controlling striatal dopaminergic function (namely, DARPP-32 and DRD2) have only been revealed in model fits to participant choices after initial learning has occurred (Frank *et al.*, 2007; Doll *et al.*, 2011). Conversely, a genetic marker of prefrontal dopamine efficacy [namely, catechol-O-methyl transferase (COMT)] related to executive function is predictive of learning rates during initial acquisition (Frank *et al.*, 2007) of model estimates of hypothesis testing and strategic exploration (Frank *et al.*, 2009a; Doll *et al.*, 2011). Similarly, functional imaging studies have shown that dopaminergic drugs modulate striatal reward prediction error signals during learning, but that these striatal signals do not influence learning rates during acquisition itself; nevertheless, they are strongly predictive of subsequent choice indices measuring the extent to which learning was sensitive to probabilistic reward contingencies (Jocham *et al.*, 2011).

In principle, with the appropriate model and task, one could separately identify integrative RL components uncontaminated from higher order prefrontal functions. Although the community has acknowledged this difficulty in identifying RL substrates, it has done so (and attempted to remedy the problem) mostly in complex learning situations, or by trying to explore specific influences on behavior (such as uncertainty). Here, we show that this mixed influence of higher order WM and lower level RL components is present in a simple instrumental learning task involving binary deterministic feedback. We show that simple RL models might miss crucial aspects of behavioral variance, or incorrectly account for that variance by attributing some observed effects to the wrong causes within the classic RL framework.

In this study, we designed a new behavioral protocol to investigate the influence of higher level cognitive systems in simple instrumental learning. We show that simple RL models cannot account for the observed behavior, and propose a hybrid model that allows us to separate the roles of capacity-limited WM from simple RL systems during learning. We further show that a genetic marker of prefrontal cortex function is associated with the WM capacity estimate of the model, whereas a genetic marker specific to BG function relates to the RL learning rate.

Materials and methods

Experimental design

All subjects gave written informed consent and the study was approved by the Brown University ethics committee. The task involved a straightforward instrumental learning paradigm in which a single stimulus was presented and subjects had to learn which of three responses to select based on binary deterministic feedback. To manipulate the WM demands separately from the RL components, we systematically varied the number of stimuli, denoted as set size n_S, to which subjects had to learn the correct actions within a block. There were six blocks in which $n_S = 2$, four blocks in which $n_S = 3$, and three blocks each of $n_S = 4$, 5, or 6 for a total of 19 blocks, and a maximum of 50 min.

Each block corresponded to a different category of visual stimulus (such as sports, fruits, places, etc.), with the stimulus category assignment to block set size counterbalanced across subjects. Block ordering was also counterbalanced within subjects to ensure an even distribution of high/low load blocks across the duration of the experiment.

At the beginning of each block, subjects were shown the entire set of stimuli and encouraged to familiarize themselves with them. They were then asked to answer as rapidly and accurately as possible. Within each block, stimuli were presented in a pseudo-randomly intermixed order, with a minimum of 9 and a maximum of 15 presentations of each stimulus, up to a performance criterion of at least four correct responses out of the five last presentations of each stimulus. As a motivational factor, subjects were given feedback indicating the number of trials (and, thus, experiment time) saved by good performance after each block.

Stimuli were presented in the center of the screen for 2 s, during which time subjects could press one of three keys. Binary deterministic auditory feedback ensued (ascending tone for correct, descending tone for incorrect), as well as a cumulating bar indicative of overall block performance. The intertrial interval was 2.5 s.

Subjects were instructed that finding the correct action for one stimulus was not informative about the correct action for another stimulus. This was enforced in the choice of correct actions, so that, for example in a block with $n_S = 3$, the correct actions for the three stimuli were not necessarily three distinct keys. This procedure was implemented to ensure independent learning of all stimuli (i.e. to prevent subjects from inferring the correct actions to stimuli based on knowing the actions for other stimuli).

Genetic sample

A total of 78 subjects (44 female) between the ages of 18 and 40 (mean 24.3 ± 5.7 years) participated in the experiment in the Laboratory for Neural Computation and Cognition at Brown University. We collected salivary DNA and investigated single nucleotide polymorphisms (SNPs) in genes associated with dopamine function, which have been previously linked to prefrontal and BG functions in learning in previous investigations (see, for review, Frank & Fossella, 2010). Specifically, we assessed the val158met SNP within the COMT gene (rs4680), an SNP within the PPP1R1B (DARPP-32) gene (rs907094), and an SNP within the DRD2 gene (rs6277). COMT is an enzyme that breaks down extracellular dopamine, with Val carriers showing more efficient COMT activity and hence lower prefrontal dopamine levels (Gogos *et al.*, 1998; Huotari *et al.*, 2002; Matsumoto *et al.*, 2003; Slifstein *et al.*, 2008). In contrast, DARPP-32 and DRD2 are far more concentrated in the striatum, where they influence dopaminergic function (Frank & Fossella, 2010). In a recent review, we identified novel non-dopaminergic candidate genes that should be investigated in RL processes due to their selective effects on direct and indirect pathway function in the BG (Frank & Fossella, 2010). We thus obtained genotype data on three of the most common SNPs associated with BG function, but which have not yet been linked to

behavior, for exploratory purposes. These included SNPs within PDYN (rs2235749), a marker of direct pathway function, and PENK1 (rs2609998) and GPR6 (rs4354185), markers of indirect pathway function. Critically, unlike DRD2, which is only preferentially expressed in the striatum but also to some extent in the frontal cortex, GPR6 is extremely specific to the striatum (see Fig. 1) (Roth *et al.*, 2006; Ernst *et al.*, 2007; Lobo *et al.*, 2007).

All subjects were successfully genotyped. Frequencies per allele were COMT 38 : 33 : 17 (Val/Val : Val/Met : Met/Met), DRD2 21 : 38 : 19 (C/C : C/T : T/T), DARPP-32 10 : 35 : 33 (C/C : C/T : T/T), PENK1 22 : 26 : 30 (T/T : C/T : TT), PDYN 13 : 34 : 31 (AA : AG : GG), and GPR6 7 : 36 : 35 (AA : AG : GG). All SNPs were in Hardy–Weinberg equilibrium (χ^2 values < 1.5, P values > 0.2), except PENK1.

The majority of the sample (51 subjects) classified themselves as Caucasian, 12 as Asian, nine as African-American, and six as 'other'. Nine individuals classified themselves as Hispanic. Because population stratification represents a potential confound for the observed genetic effects, several additional measures were taken to verify that the effects reported herein were not due to admixture in the sample. Allele frequencies did not differ from Hardy–Weinberg equilibrium in any subgroup when analyzed independently. There was little evidence to suggest that the genetic effects observed in the present study were due to population admixture in the sample (further tested in Results).

Genotyping method

The DNA was collected via 2 mL salivettes (DNA Genotek Oragene). Samples were genotyped by the Mind Research Network (Albuquerque, NM, USA) using TaqMan primer and probe pairs; the probes were conjugated to two different dyes (one for each allelic variant). Taqman assays were designed and selected using the SNPBrowser program (Applied Biosystems). The polymerase chain reaction mixture consisted of 20 ng of genomic DNA, 1× Universal PCR Master Mix, a 900 nM concentration of each primer, and a 200 nM concentration of each probe in a 15 μL reaction volume. Amplification was performed using the TaqMan Universal Thermal Cycling Protocol, and fluorescence intensity was measured using the ABI Prism 7500 Real-Time PCR System. Genotypes were acquired using the 7500 system's allelic discrimination software (SDS version 1.2.3).

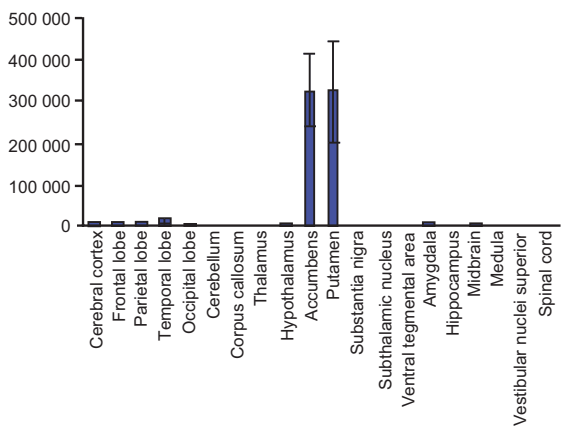

FIG. 1. GPR6 expression is specific to the striatum. Relative expression of GPR6 for individual tissues in postmortem human brain. Note the highly specific expression in the striatum (accumbens and putamen). Reproduced from Roth *et al.* (2006).

Results

Behavioral analysis

Overall, performance was quite good (see Fig. 2A); for all set sizes, the final accuracy (defined as the mean performance over the last two trials of each stimulus) was > 94%, and the time to asymptote (defined as mean number of stimulus presentations per block) was < 11. As expected, learning varied as a function of set size, in terms of both trials to asymptote and mean performance over the whole block ($r = 0.3$ and $r = -0.53$, both P values < 10^{-4}). The final asymptotic accuracy was high for all set sizes (ranging from 94% to 97%).

We next sought to investigate the source of the decrease in learning performance linked to larger set sizes. We identified two potentially separable causes: WM load and delay. With limited WM capacity, subjects might only be able to maintain the response–outcome associations for a subset of the stimuli; this limitation would be more evident for larger set sizes as capacity was exceeded. Furthermore, even if a given stimulus association was updated into WM, the delay until that same stimulus was next observed was on average longer for larger set sizes, thus increasing the opportunity for memories to be degraded due to decay or due to having updated other stimuli in the interim, given limited resources. Although memory load and average delay were both highly correlated with set size, these two factors can be studied orthogonally in our experimental design. Indeed, although memory load was fixed within a block, delay was locally variable within each block.

To investigate this issue, we assessed the degree to which memory deviated from optimal as a function of delay. We thus restricted our analysis to trials for which the subject had already responded correctly to the given stimulus at least once, such that any deviation from optimality thereafter reflected degraded memory (given the deterministic reinforcement contingencies). As expected, performance decreased with increasing delay (Fig. 2B and C), especially with increasing set size. This set-size effect was not solely due to larger delays for higher set sizes, e.g. the same pattern observed for a delay of > 2 was observed with a delay of exactly 3 (data not shown).

In addition to these robust effects of delays and set sizes, we also observed a surprising negative effect of set size for delay 1, which, although more subtle, was nevertheless significant (see Fig. 2B). This indicated that, when the same stimulus was presented twice in a row, subjects were more likely to make an error in the second trial after having just responded correctly to that stimulus for lower set sizes. This finding may reflect a lower degree of task engagement for easier blocks, leading to a slightly higher likelihood of attentional lapses. Indeed, repetition errors were overall faster than correct repetitions ($t = 4.4$, $P < 10^{-4}$), and this effect was greater for lower set sizes ($P = 0.05$), indicating a faster/less accurate speed–accuracy trade-off in these blocks.

In order to analyze the different effects in a more principled way, we employed logistic regression. We analyzed the proportion of correct responses on these trials as a function of three variables.

1. The set size, determining the overall load in which this trial was embedded.
2. The delay in the number of trials since the subject correctly responded to the current trial's stimulus.
3. The total number of previous correct responses for the current trial's stimulus.

We performed this logistic regression within each subject, and then assessed significance across the whole group. This analysis allowed us to test whether there were effects of set size independent of delay and vice versa. There was a main effect of set size ($t = -6.6$, $P < 10^{-4}$) and a main effect of number of correct repetitions ($t = 14$, $P < 10^{-4}$), with

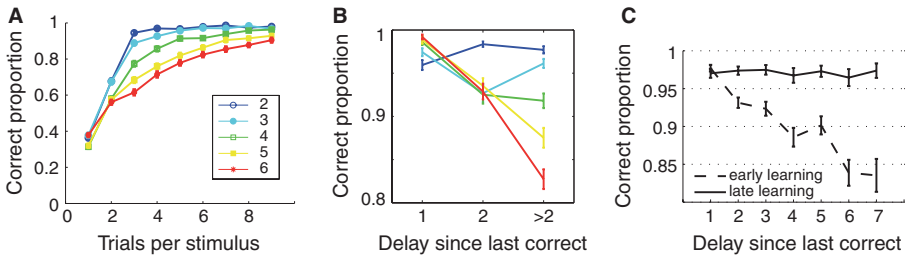

FIG. 2. Behavioral results. (A) Proportion of correct responses as a function of the number of times that each stimulus was presented, plotted separately for each set size. Error bars reflect SEM. (B) Proportion of correct responses as a function of the trial interval since the current stimulus was last reinforced (delay), separately for each set size. (C) Proportion of correct responses as a function of the delay, for early learning trials (one to three correct responses for the current stimulus) or late learning trials (seven or more correct responses). The effect of delay disappears over learning.

performance advantages for both decreasing set sizes and increasing number of correct repetitions. Although the negative effect of delay was not significant, delay interacted with the number of correct repetitions ($t = 1.9$, $P = 0.06$), indicating that it was detrimental to performance during early learning, with a diminution of this effect over time (see Fig. 2C). The set-size effect also interacted with the number of correct repetitions ($t = 4.4$, $P < 10^{-4}$), such that the effect of set size decreased with learning. These results support the notion that, with higher set sizes as WM capacity was exceeded, subjects relied on more incremental RL, and less on delay-sensitive memory.

The predicted probabilities from the logistic regression were qualitatively similar to the observed learning curves (see Fig. 3, top), showing that these three predictors captured a substantial degree of behavioral variance. Importantly, the logistic regression also allowed us to investigate the effects of a single factor controlling for the other factor, and to replot the corrected learning curves. We thus computed predicted probabilities fixing the delay to the minimal value of 1, to generate delay-corrected learning curves (Fig. 3, bottom left). Conversely, we fixed the set size to the minimal value of 2 to generate set-size-corrected learning curves (Fig. 3, bottom right). As expected from the significant effect of both predictors, the block set-size effect remained in corrected learning curves, showing that neither predictor alone can account for slower learning in higher load blocks. Note that in the set-size-corrected learning curves there remained differential effects of set size, due to the larger delays with increasing set size.

Next, we examined multiple computational models that tested the relative influences of WM and RL mechanisms and their interactions. We focus here on five alternative models. As a baseline, we introduce a basic RL model. The other four models include an RL model with variable learning rate, an RL model with forgetting, a pure WM model, and a model incorporating both WM and RL.

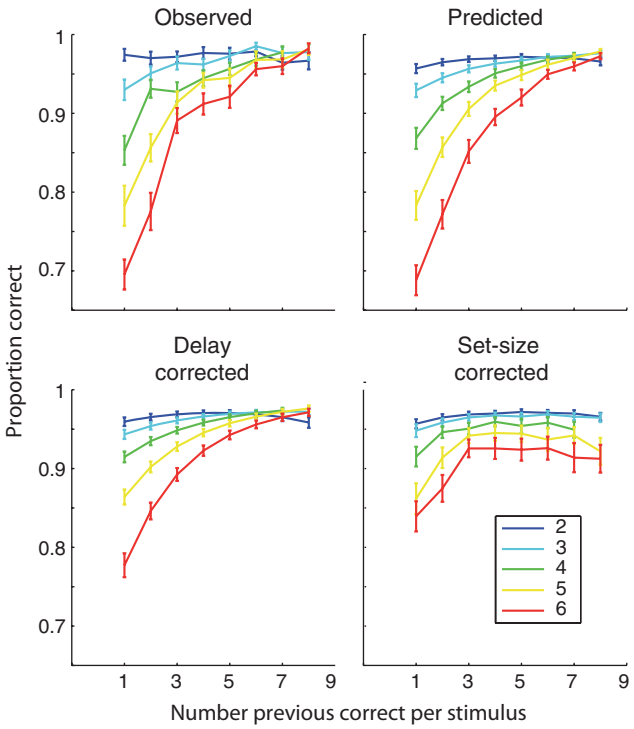

FIG. 3. Logistic regression. Top left: Observed proportion of correct responding for a given stimulus as a function of the number of previous correct trials for that stimulus, separately for each set size. Top right: Predicted probability from logistic regression controlling for other factors (delay since last stimulus-specific correct trial, set size, and their interactions). Bottom: Corrected learning curves, plotting predicted probabilities from logistic regression with one predictor value artificially fixed. Bottom left: Delay-corrected learning curve (delay predictor fixed to minimum delay). Bottom right: Set-size-corrected learning curve (set size predictor fixed to minimum set size).

Models

Pure reinforcement learning model, two parameters (RL2)

We begin with the standard RL model, in which for each stimulus (or state), s, and action, a, the expected reward $Q(s, a)$ is learned as a function of reinforcement history. Specifically, the Q value for the selected action given the stimulus is updated upon observing each trial's reward outcome, r_t, as a function of the prediction error between expected and observed reward

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_{\mathrm{RL}} \times (r_t - Q(s_t, a_t)) \qquad (1)$$

where α_{RL} is the learning rate. Choices are generated probabilistically as a function of the difference in Q values between the available actions using the softmax choice rule:

$$\mathrm{p(a|s)} = \frac{\exp(\beta_{\mathrm{RL}} Q(s, a))}{\sum_i \exp(\beta_{\mathrm{RL}} Q(s, a_i))} \qquad (2)$$

where β_{RL} is an inverse temperature determining the degree to which differences in Q values are translated into a more deterministic choice.

In this two-parameter model, learning for each stimulus is determined only by the sequence of actions and rewards for that specific stimulus. In particular, it is unaffected by the relative sequence of other stimuli, delay, or the number of stimuli in the set (see Fig. 4). Thus, this model cannot predict any effect of set size or delay. It serves as a benchmark to evaluate other models and to control for general

effects of learning across the block. It also corresponds to the generic corticostriatal learning model.

Multiple learning rate reinforcement learning model (RL6 and RL10)

To account for different learning curves in an RL framework, we simply allow the learning rate α_{RL} to vary as a function of set size. This model is of course more complex (having six parameters), but allows us to assess whether learning rates vary systematically by set size. We also tried various variants of this model, including a 5 β single learning rate model and even a 5 β, 5 α model (10 parameters). Because most experiments only include one set size, this last 10-parameter model effectively treats different set sizes as different experiments.

Forgetful reinforcement learning (RLF) model

In this model, Q learning and action selection occur as in Eqns 1 and 2 from model RL2. Additionally, we include a supplementary effect of forgetting across time; at each trial, for all stimulus–action pairs (s, a), we decay Q values towards their initial values

$$Q(s,a) \leftarrow Q(s,a) + \epsilon \times (Q_0 - Q(s,a)) \tag{3}$$

where $Q_0 = 1/n_A$ is the initial Q value for all actions, representing random policy, and ϵ controls the degree of forgetfulness.

Thus, with increasing delay between repeated encounters with the same stimulus, the more the learned Q values will have decayed. Consequently, with $\epsilon > 0$, this three-parameter model predicts a decrease in performance in higher set-size blocks, due solely to the increased average delay between stimulus repetitions for higher set sizes (see Fig. 4).

Reinforcement learning + working memory model

In the RL + WM model, action selection derives from a mixture of a pure simple RL model (i.e. RL2) applied to all set sizes, together with a limited capacity WM component. We simulated WM as the encoding of an observed event that, if maintained in memory, could serve to immediately and robustly affect behavior. That is, perfect memory could be represented by a Q learning system with a learning rate of 1 (which is optimal for a deterministic task). However, memory degrades over time and is capacity-limited. For the degradation effect, we implement a decay as in the RLF model so that, after RL update, for all (s, a) at each trial

$$Q_{\mathrm{WM}}(s,a) \leftarrow Q_{\mathrm{WM}}(s,a) + \epsilon \times \left(\frac{1}{n_A} - Q_{\mathrm{WM}}(s,a)\right) \tag{4}$$

The probability of action selection according to the WM component is then $p_{\mathrm{WM}}(a) = \mathrm{softmax}(\beta_{\mathrm{WM}} Q_{\mathrm{WM}})$. As in earlier models, the probability of action selection for the RL component is $p_{\mathrm{RL}}(a) = \mathrm{softmax}(\beta_{\mathrm{RL}} Q_{\mathrm{RL}})$. As stated thus far, the WM component captures forgetting but does not yet account for the known limited capacity of WM. This capacity limitation is factored into the mixture weight $w(t)$ determining the probability that action selection is governed by the RL or WM component

$$p(a) = (1 - w(t)) p_{\mathrm{RL}}(a) + w(t) p_{\mathrm{WM}}(a) \tag{5}$$

Crucially, the mixing parameter $w(t)$ varies as a function of time and is dependent on block set size relative to individual capacity size C. In particular, if set size $n_S \leq C$, all stimuli can be remembered deterministically, whereas if set size $n_S > C$, a given stimulus can only be probabilistically stored in WM, with probability C/n_S. Thus, the likelihood p_{WM} has to be adapted to reflect this probability. Specifically, the likelihood p_{WMC} that the capacity-limited WM component would correctly predict the current reward observation r_t is

$$\begin{aligned} \text{if } r_t = 1,\ p_{\mathrm{WMC}}(r_t \mid s_t, a_t) &= \min(1, C/n_S) \times Q_{\mathrm{WM}}(s_t, a+t) \\ &\quad + (1 - \min(1, C/n_S)) \times 1/n_A \\ \text{if } r_t = 0,\ p_{\mathrm{WMC}}(r_t \mid s_t, a_t) &= \min(1, C/n_S) \times (1 - Q_{\mathrm{WM}}(s_t, a_t)) \\ &\quad + (1 - \min(1, C/n_S)) \times 1/n_A \end{aligned} \tag{6}$$

In words, the likelihood varies directly with the relative capacity to set-size ratio, such that, when capacity is exceeded ($C/n_S < 1$), the WM component is less likely to correctly predict the reward, with increasing contribution of randomness. In contrast, the likelihood term

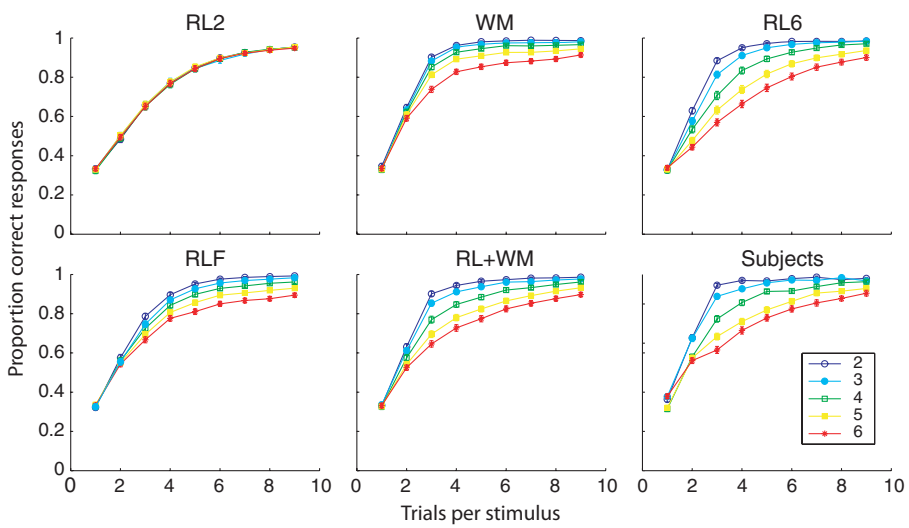

FIG. 4. Model results. Learning curves as a function of set size for each model, generated using best-fit parameters for each subject given the model. Models: RL2, two-parameter RL model; WM, pure WM model; RL6, RL model with five learning rates + one softmax temperature; RLF, three-parameter forgetful model; RL + WM, mixture RL and WM model. Subjects, observed learning curves across all subjects (from Fig. 2).

corresponding to the RL model p_{RL} is not capacity-limited and is computed in a straightforward manner from the RL Q values

$$\begin{aligned} &\text{if } r_t = 1,\ p_{RL}(r_t \mid s_t, a_t) = Q(s_t, a_t) \\ &\text{if } r_t = 0,\ p_{RL}(r_t \mid s_t, a_t) = 1 - Q(s_t, a_t) \end{aligned} \tag{7}$$

Given these two likelihoods, p_{WMC} and p_{RL}, the relative reliability of the WM compared with the RL system is then inferred over time to influence the mixture weight according to a Bayesian model averaging scheme

$$w_{n_s}(t+1, s) = \frac{p_{WMC}(r_t \mid s_t, a_t) w_{n_s}(t, s)}{p_{WMC}(r_t \mid s_t, a_t) w_{n_s}(t, s) + p_{RL}(r_t \mid s_t, a_t)(1 - w_{n_s})} \tag{8}$$

Thus, with high confidence in the WM system (when set size is within capacity), action selection is primarily determined by WM, whereas as capacity is exceeded there is an increasing contribution of RL. Initial mixture weights are initialized to represent initial confidence in WM efficiency, again assuming a fixed C

$$w_{n_s}(t = 0, s) = w_0 \times \min\left(1, \frac{c}{n_s}\right) \tag{9}$$

This initialization reflects the fact that participants are more likely to begin with a high WM contribution when they are shown stimulus sets with low set size.

This six-parameter model (β_{RL}, α_{RL}, C, β_{WM}, w_0, ϵ) can simulate both delay and set-size effects independently, due to forgetting and capacity, respectively. It also predicts that, whereas WM should be a better predictor of outcomes initially (due to faster learning rate), over time, RL should supersede WM. Indeed, incremental accumulation of RL values should perfectly predict the outcomes, whereas WM is still subjected to decay and capacity limits. This corresponds to automated or habitual choice, and predicts a reduction of the frequency effect over time. Note that it also involves only a single RL process that does not differ between set sizes, and has no forgetting.

Pure working memory model (no reinforcement learning)

As a control, we also fitted and simulated a pure WM model. This model corresponds to the WM part of the previous model, mixed with random action selection when capacity is exceeded (i.e. no RL component). Mixture weight is then fixed to the initial mixture weights in the previous model. This model may predict both set-size and delay effects (see Fig. 4).

Model comparisons

For all models, we used the standard fitting procedure of selecting model parameters that maximized the likelihood of observed action choices conditioned on the history of observations and the model parameters. We verified that the parameters of the full RL + WM model are identifiable by generating data from this model for various parameter values, and then recovering these values to a satisfying degree of precision. Although we focus our analysis using maximum likelihood estimates for each individual participant, we also confirmed that the results reported below hold for a simple hierarchical fitting procedure in which summary statistics for each parameter were estimated across the entire group of subjects and which then acted as priors for the estimation of individual subject parameters (Daw, 2011).

For absolute fit measures, we first present pseudo-r^2 values, which scale the log likelihood in comparison to that of a random model selection. However, as this is a deterministic learning task, with relatively good performance, any model predicting an overall increase in accuracy with time will capture much of the variance relative to chance. Thus, we used the RL2 model as a baseline, and rescaled log-likelihood measures for each model relative to this baseline

$$\Delta(\text{model}) = \frac{\text{LLH(model)} - \text{LLH(RL2)}}{\text{LLH(RL2)}} \tag{10}$$

For model comparison, we penalized more complex models (with more parameters) by computing the Akaike information criterion (AIC) for each subject. We also report exceedance probabilities reflecting the likelihood that each model is the best of all candidate models given the distribution of AIC values across all subjects (Stephan *et al.*, 2009). (We also verified that the AIC statistic more appropriately reflects model fit than the alternative Bayesian information criterion. In particular, simulations confirmed the often observed result that the Bayesian information criterion overpenalizes model complexity; when generating data from the RL + WM model, AIC correctly identified this model as the best fit to the generated data, whereas the Bayesian information criterion statistic overpenalized this model.)

We then generated simulated data from each model using the best-fit parameters for each subject from each model. This procedure allows us to determine whether the model fits capture key qualitative features of the data. Twenty simulated experiments were generated with each subject's individual sequence of stimuli and parameters, and then averaged to represent that subject's contribution. Choice probabilities were then averaged across subjects (see Fig. 4). Similar results were obtained when simulating models with median fitted parameters across the group.

Model results

All models afforded good fits to the data (minimum average pseudo-r^2 was for RL2 at 0.55), and all but the simpler model RL2 reproduced qualitatively the observed set-size effect on behavior (see Fig. 4).

Note first that the RL + WM model accounts best for subjects' data across all criteria (significantly better AIC than all other models, exceedance probability of 1.0 across the whole group, and high proportion of subjects best fitted by RL + WM; see Fig. 5). Thus, accounting for both capacity-limited WM and RL provides a better fit of the data than either process on its own (i.e. pure WM or pure RL models). Importantly, there was no trade-off in estimated parameters between the two main parameters of interest: capacity and RL learning rate, as would be revealed by a negative correlation between them. In addition to these fit measures, a few interesting points should be noted about the best-fitting parameters of various models.

First, RL6 can provide a better qualitative fit to the data than RL2, by allowing RL parameters to vary independently across set sizes and hence capture the differential learning. Notably, the fitted learning rates α in RL6 systematically decreased with set size, thus accounting for slower learning in higher load learning situations. Note, however, that in addition to the fact that this model is clearly not parsimonious, it is also not sufficient to account for all set-size effects; even without penalizing for model complexity, it did not provide a better fit than the RL + WM model, which has a single set of RL parameters. This is also true of the model RL10, which effectively treats different set sizes

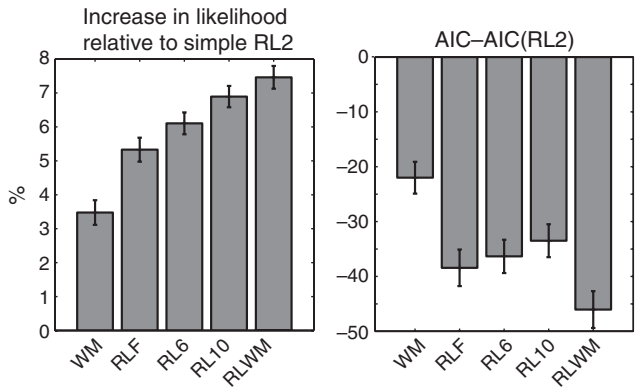

FIG. 5. Model fits. Left: Relative increase in likelihood compared with simple RL2 (see Eqn 10). All pairwise comparisons are significant. Right: Difference in model's AIC compared with baseline (negative values indicate better fit penalizing for additional parameters). All pairwise comparisons are significant. Models: RL2, two-parameter RL model; WM, pure WM model; RL6, RL model with five learning rates + one softmax temperature; RLF, three-parameter forgetful model; RL10, RL model with five learning rates + five softmax beta parameters; RLWM, mixture RL and WM model.

as different experiments modeled with different RL parameter pairs (α, β) and thus has 10 parameters: even so, this models provides significantly worse likelihood than the RL + WM model, showing that it is missing crucial aspects of the variance (see Fig. 5).

Second, adding the forgetting parameter to the basic RL model (i.e. RLF) improved the fit significantly, showing that an important part of set-size effects could be accounted for by taking delay effects into account. Although delay effects are more germane to the WM than the RL system, this finding is particularly relevant for other experiments where only one set size is used, and suggests that some behavioral variance due to WM can be accounted for by incorporating this simple forgetting mechanism. Indeed, the best fitting softmax inverse temperature parameter β was significantly greater for RLF than RL2, probably because lower values are needed in RL2 to account for errors, whereas in RLF these are largely explained by an explicit forgetting mechanism (Table 1).

Interestingly, the best-fitting RL + WM model yielded parameter estimates for capacity with average values between 3 and 4 (mean $C = 3.7 \pm 0.14$, Table 1). This is within the expected bounds of human WM capacity. It also best captures the qualitative learning pattern (see Fig. 4), despite a slight underestimation (present equally in all models) of initial performance in low load blocks ($n_S = 2$ and $n_S = 3$). We hypothesize that this might be due to not taking into account an increase in WM performance when set size is under capacity, thus affording possible redundant encoding of information [e.g. the slots + averaging model of Zhang & Luck (2008)]. However, additional parameters would be needed to estimate this effect and more work is needed to clarify this point.

Genetic results

We investigated the hypothesis that WM effects play a strong role in RL experiments by assessing the effect of prefrontal dopamine efficiency on performance, through the val158met polymorphism in the COMT gene (Egan *et al.*, 2001; Tunbridge *et al.*, 2004; Frank *et al.*, 2007, 2009b; Slifstein *et al.*, 2008; de Frias *et al.*, 2010). In accordance with the existing literature, we hypothesized that carriers of the Val allele would have less efficient WM, and thus would show specific reductions in aspects of learning that depend on WM compared with Met/Met homozygous subjects.

Importantly, the RL + WM model provided the best fit to the data for both groups across all fit criteria, with an exceedance probability no lower than 0.997 (although fit values overall were larger for the Met/Met group, due to overall better performance), allowing us to interpret parameter differences for this model between the groups. Notably, there was a significant effect of COMT genotype on the WM capacity parameter (Fig. 6D), with significantly greater capacity for Met/Met ($N = 17$) than Val ($N = 61$; Wilkoxon $z = 2.5$, $P = 0.012$). There was no group difference for any other model parameter. This effect also did not interact with ethnicity ($F = 0.64$, ns), and indeed the mean capacity values for each genotype were nearly identical in Caucasians alone and in the smaller non-Caucasian sample.

We next investigated the origin of this capacity effect in the behavior. Figure 6 (top) shows behavioral learning curves for the Met/Met group and Val group. Although both groups learned to similar asymptotic levels in all block set-size conditions, initial learning was specifically slower for Val in higher load conditions. To further investigate this genetic effect, we concentrated on performance in the third to fifth presentation of each stimulus as the most representative of learning speed (because subjects require a minimum of two trials of experience to be able to know the correct motor response, given that there are three possible responses). Indeed, optimal learners would be perfect on these 'middle trials'. The middle performance in high load blocks ($n_S = 5, 6$; Fig. 6C) was significantly lower for Val carriers compared with the Met/Met group ($t = 2.03$, $P = 0.045$), which was not the case for low load blocks ($t = 0.92$, ns). The interaction between COMT genotype and load condition (high vs. low) on these middle trial performances approached significance ($t = 1.86$, $P = 0.067$).

We also searched for neural correlates of the slow but integrative accumulation of evidence represented in the RL part of the RL + WM model. We investigated polymorphisms in genes that have previously been found to be implicated in such learning, and associated with striatal function in the Go and No-Go pathways, DARPP-32 and DRD2, respectively (see e.g. Frank *et al.*, 2007, 2009a; Doll *et al.*, 2011), with the hypotheses that they might affect the model parameters specifically related to RL (β_{BG}, α_{BG}). However, we found no effects of these genes on these parameters (or any other parameter).

In a somewhat more exploratory but still hypothesis-driven analysis, we investigated variants in three non-dopaminergic genes that have been shown to be specifically expressed in the BG and can thus serve as indexes of BG function. The three polymorphisms analyzed were selected from those identified by Frank & Fossella (2010) as candidates for the analysis of RL given their specificity to striatal function. In particular, GPR6 is very specifically expressed in the human striatum, including the caudate, putamen and nucleus accumbens and not in the cortex or any other area (see Fig. 1) (Roth *et al.*, 2006; Ernst *et al.*, 2007). Although the impact of this gene on human behavior has not yet been investigated, GPR6 has been implicated in instrumental learning in mice (where it is also striatal-specific). In particular, knock-out of GPR6 resulted in improvements in BG-dependent instrumental learning (Lobo *et al.*, 2007) without having any other phenotypes. We thus had a strong *a-priori* hypothesis that this gene would impact the BG-specific part of the model. Nevertheless, because there are not previous reports of GPR6 effects on human behavior, we considered this analysis to be exploratory together with that of two other striatal polymorphisms. Thus, for significance testing, we corrected for multiple comparisons across these three exploratory SNPs.

Interestingly, we found a strong gene dose effect on the RL learning rate parameter α_{BG} (see Fig. 6E, $r = 0.33$, $P = 0.002$), which was also significant in a non-parametric analysis comparing GG with A carriers

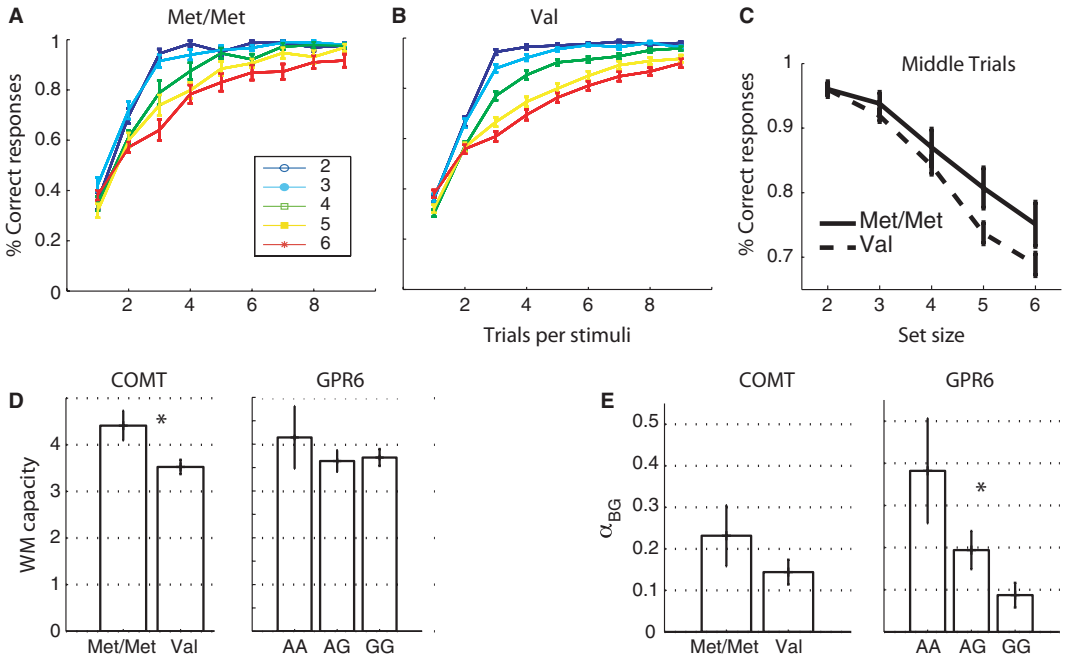

FIG. 6. Gene effects. Top: COMT behavioral effects. (A) Learning curves for homozygous Met carriers and (B) Val carriers. (C) Proportion of correct responses in middle trials as a function of genotype and set size. Val carriers exhibited a specific impairment in high load blocks compared with Met/Met for middle trials (but not for later learning trials as the WM component is superseded by RL). (D) Fitted capacity (C) parameter from the RL + WM model for COMT and GPR6 genotypes. (E) Fitted BG learning rate parameter α_{BG} from the RL + WM model for COMT and GPR6 genotypes. * indicates a significant gene effect ($P < 0.05$).

($t = 2.5$, $P = 0.015$). This effect did not interact with ethnicity ($F = 1.26$, ns) and remained significant in the group of 51 Caucasians ($t = 2.6$, $P = 0.011$). There was no effect of GPR6 on any other model parameters, and no other gene effect survived significance after correction for multiple comparison. In particular, as a *post hoc* comparison, we checked that there was no effect of GPR6 on the WM specific model parameter capacity C ($F = 0.18$, $P = 0.67$).

Given this highly specific effect of GPR6 on the RL learning rate, we reasoned that this effect should be observable in behavioral analysis. In particular, the contribution of the incremental RL system should have the greatest effect in late learning trials, due to more learning, and especially for high load (for which it would supersede WM due to capacity limits). Indeed, relative to GG homozygotes, A carriers showed specific impairments in high vs. low load blocks for late learning trials ($t = 2.17$, $P = 0.03$), but not early learning trials ($t = 0.9$, ns; although there was a global impairment irrespective of load during early learning; $P = 0.027$). Note that this is the opposite pattern to that found for the COMT polymorphism, where there was an interaction effect with load when WM was more prevalent on early trials, and no interaction on late trials ($t = 1.2$, ns).

There was also no correlation between allelic expression in GPR6 and COMT genotypes ($r = 0.05$, $P = 0.54$), thus indicating that the genetic effects are independent. Taken together, these results show a double dissociation between the genes indexing prefrontal cortex and basal ganglia function (COMT and GPR6) on model parameters indicating WM vs. RL influence on learning behavior.

Discussion

Human learning involves a network of cognitive processes and a multitude of neural mechanisms. The vast majority of neuroscientific studies of RL have focused on mechanisms underlying the class of 'model-free' RL algorithms; they capture the incremental learning of values associated with states and actions, without considering the extent to which the subject (human, animal, or computer model) can explicitly plan which actions to make based on their knowledge about the structure of the environment. Although the use of these simple model-free algorithms, and mechanistic implementations thereof in BG circuits, has been extremely profitable, it is also acknowledged that they only capture a component of instrumental learning. Indeed, another class of 'model-based' RL algorithms considers explicitly the subject's tendency to build a world model.

Many types of world model learning have been investigated, including learning of sequential hierarchies (Botvinick *et al.*, 2009;

TABLE 1. Model parameters

	β_{BG}	α_{BG}	α_2	α_3	α_4	α_5	α_6	ϵ	C	w_0	β_{WM}
RL2	7.1 (0.6)	0.38 (0.02)	–	–	–	–	–	–	–	–	–
WM	–	–	–	–	–	–	–	0.1 (0.005)	5.4 (0.8)	0.95 (0.005)	12 (1.3)
RL6	7.8 (0.6)	–	0.86 (0.02)	0.68 (0.03)	0.45 (0.03)	0.32 (0.02)	0.22 (0.02)	–	–	–	–
RLF	24.1 (1.9)	0.29 (0.02)	–	–	–	–	–	0.07 (0.003)	–	–	–
RL + WM	26.6 (3.3)	0.16 (0.03)	–	–	–	–	–	0.23 (0.02)	3.7 (0.14)	0.81 (0.02)	45 (4.5)

Mean (SEM) fitted parameters for five representative models. Models: RL2, two-parameter RL model; WM, pure WM model; RL6, five-learning-rate RL model; RLF, three-parameter forgetful model; RL + WM, mixture RL and WM model.

Daw *et al.*, 2011; Ribas-Fernandes *et al.*, 2011), environment volatility (Behrens *et al.*, 2007), and abstract structure (Doya, 2002; Hampton *et al.*, 2006; Imamizu *et al.*, 2007; Frank & Badre, 2011, Collins & Koechlin, unpublished data). Some of these works also propose arbitrations between different parallel ways of using model information (e.g. see Lengyel & Dayan, 2007 for a three-controller system). A popular version of 'model-based' RL concentrates on the construction of a world model that learns sequential dependencies, in terms of transition probabilities, from one state to the next and how they are affected by the agent's actions. Planning is implemented by searching a tree of these transition probabilities (e.g. in forward direction from the current state to arbitrary depths of anticipated future states/actions) to select the path with the highest expected value. Theoretical work has suggested the existence of parallel striatal model-free and prefrontal model-based controllers (Daw *et al.*, 2005). Notably, these authors explicitly consider the notion that, although the model-based system is more flexible, it also affords additional computational cost associated with searching the tree, and eventually gives way to a model-free system as state-action values become well learned.

It is clear that one such computational cost is the capacity limitation of WM, which would be required to maintain a set of if/then relationships in mind in order to plan effectively. Indeed, this distinction between model-free and model-based control is closely related to a similar distinction in the cognitive psychology of dual systems linked to automatic, associative processing vs. controlled, deliberative but capacity-limited processing (Evans, 2003; Sloman & Hagmayer, 2006). This secondary process is often attributed to the prefrontal cortex and its involvement in WM.

In this study, we have shown that these higher level processes, including capacity-limited WM with decay, are crucially involved in even simple learning tasks typically studied with traditional model-free RL algorithms, in such a way that is not accommodated by these algorithms. Indeed, we showed that, if such algorithms are used, the parameters necessarily adjust to accommodate behavioral variance, but then these parameters are no longer estimates of the intended RL processes and are therefore misleading. Even when separate RL parameters were estimated for each set size in our experiment (an implausible, non-parsimonious model with 10 parameters), it did not provide as good a fit to the data as did our simpler hybrid model estimating WM contributions together with a simple process. Moreover, we showed separable individual differences in learning that are attributed to genetic variance in prefrontal and striatal mechanisms.

The experimental protocol allowed us to determine that variance in behavior was explained separately by two characteristics of WM: its capacity and its stability. Indeed, behaviorally, learning was slower in problems with greater load, but there were minimal differences in asymptotic performance. Furthermore, although performance was initially highly subject to degradation due to the delay since the presented stimulus was last observed, this delay effect disappeared over learning.

We designed a new mixture model allowing us to separately represent contributions of the RL–BG system and WM–prefrontal cortex system, as well as the dynamic allocation of responsibility to each in action selection. The RL part of the model is the simplest version of the widely used *Q*-learning algorithm, where action values for each stimulus are learned independently from those of other stimuli, and are insensitive to delay or capacity. The WM part of the model is sensitive to load effects, and representations of observed events decay over time. Further, the dynamic allocation of action selection to the different systems takes into account the limited capacity of the WM system and the relative reliability of each system over time. As such, with increasing experience, the RL system accumulates sufficient evidence and, because of its lack of capacity limitations or memory decay, it eventually supersedes the WM system. This point is reminiscent of that of Daw *et al.* (2005), but here the arbitration between the RL and WM systems is determined by direct estimates of their relative values or reliabilities, rather than the uncertainties about them.

It is of interest to note the crucial role played in confronting qualitative predictions from models with key features of the behavior for identifying valid models. Indeed, although quantitative measures of fit, such as likelihood and the AIC, are essential to inform on the predictive power of a model trial-by-trial, they can be somewhat misleading by themselves. Firstly, they are most often interpreted as measures of relative fit in comparison to other models, rather than absolute fit, as absolute values such as pseudo-r^2 depend heavily on experimental design. Secondly, penalization of parameter complexity can be imperfect and lead to inappropriate model selection. It is therefore essential to demonstrate that the model used to fit the data can reproduce the behavioral effects of interest, here most evident in terms of effects of set size and delay.

In addition to capturing the key features of the behavioral data with relatively few parameters, this RL + WM model also allowed us to separately estimate WM capacity and RL learning rates. Notably, the capacity estimate ranged between 3 and 4 for most subjects, coherent with the widely accepted values in the existing literature (Cowan, 2010). Furthermore, having factored out these WM contributions, we also obtained a more reasonable estimate of a single RL learning rate (a value of 0.16, rather than the much higher values obtained, especially for lower set sizes, in the pure RL models). Finally, individual differences in WM capacity and RL learning rate were separately predicted by COMT and GPR6 genotypes.

The current exercise also serves to illustrate that the quality and interpretability of model fits depend not just on the suitability of the model itself, but on the task conditions manipulated by the experiment. Most RL experiments do not vary the stimulus set size, so that (again regardless of whether WM is built into the model) the load aspect of WM influence on learning would not be easily observed. However, many experiments do include randomized sequences of multiple stimuli, and in these cases the effects of variable delay and time-dependent WM effects could potentially be observed. Indeed, we found that the simple RLF model, which includes a simple time-decay mechanism in a basic RL framework, accounts for significantly more variance than RL2. Thus, accounting for this decay mechanism should significantly improve model fit to human instrumental learning experiments. However, two problems will remain. Firstly, this decay is fixed over trials, whereas we showed here that these WM effects disappear as RL learning progresses. Secondly, this RLF formulation includes in a single value system the effects of different neural mechanisms, thus making it more difficult to dissociate their actual correlates. As an example, had we only considered the basic RLF and RL2 models, we would have concluded that the RLF model provided the best fit to the data and then would be justified in analyzing its parameters. *Post hoc* inspection of these parameters revealed strong effects of COMT on all three RLF parameters (such that one might have incorrectly attributed COMT effects in modulating the RL learning rate), and no effect of GPR6 on any of them (as if it did not affect learning rate).

Instead, devising a model that dissociates the relative influences of RL and WM on learning behavior has allowed us to exhibit a double dissociation on the neural correlates of instrumental learning. COMT polymorphisms selectively predicted differences in the WM capacity of subjects, as exhibited in a selective effect on performance in higher load problems, and GPR6 polymorphism selectively predicted differences in the RL learning rate. No other effects on other parameters of

these two genes were observed. The model thus successfully allowed us to dissociate separate influences on learning from separate neural systems, in separate components of the model. Of course, we acknowledge that the same logic brought forth in the preceding paragraph may also apply to our analysis; perhaps if we had developed a yet better fitting model, we might have derived different conclusions about the neurogenetic components. In fact, this is a problem that pervades all of science, not just RL model fitting. As such, our assignment of neurogenetic effects on specific model parameters awaits further replication across other tasks, manipulations, etc. Nevertheless, it is important to emphasize that we also found clear behavioral correlates for these effects that are independent of the specific model (COMT on accuracy in early learning trials with increasing load, and GPR6 on general learning speed).

The effects of the val158met COMT polymorphism have been widely studied in the executive function literature, showing better WM, attention or goal-directed performance for homozygous met allele carriers (Egan *et al.*, 2001; Goldberg & Weinberger, 2004; Blasi *et al.*, 2005; Bruder *et al.*, 2005; Frank *et al.*, 2007, 2009a; Tan *et al.*, 2007; Green *et al.*, 2008; Doll *et al.*, 2011). A recent meta-analysis showed that the COMT genotype reliably affects prefrontal activation (effect size $d = 0.73$), with an advantage for met carriers in executive function (Mier *et al.*, 2010). Indeed, the COMT gene codes for the COMT enzyme that degrades extracellular dopamine. Met-allele versions of this enzyme are less efficient in degrading dopamine, thereby enabling sustained dopamine levels to persist in the PFC, and promoting the stability of actively encoded WM representations (Durstewitz & Seamans, 2008; Durstewitz *et al.*, 2010). Although COMT is also present in the BG, the effects of its manipulation appear to be relatively negligible, due to the presence of much more efficient active dopamine transporters for reuptake (Gogos *et al.*, 1998; Sesack *et al.*, 1998; Huotari *et al.*, 2002; Matsumoto *et al.*, 2003; Tunbridge *et al.*, 2004). Consistent with these findings, we found here that COMT homozygous met-allele carriers exhibited significantly higher estimated WM capacity than did val carriers. Note that the effect was found specifically on the discrete capacity limitation parameter (Zhang & Luck, 2011). Other model parameters implemented more continuous aspects of WM limitations, through precision or temporal stability (Bays & Husain, 2008), but the COMT effects were selective to the discrete capacity.

In stark contrast to the extensive literature on COMT, to our knowledge there are no existing human behavioral studies examining GPR6. However, because this gene is very specifically expressed in the BG, in the indirect pathway in particular (Roth *et al.*, 2006; Ernst *et al.*, 2007; Lobo *et al.*, 2007), and because GPR6 mutations affect instrumental conditioning in rodents (Lobo *et al.*, 2007), the current GPR6 SNP was recently identified as a strong candidate for the non-dopaminergic investigation of BG-specific RL processes in humans (Frank & Fossella, 2010). In this study, GPR6 polymorphism was strongly correlated to the BG learning rate parameter of the model, thus relating directly this general, widespread, slow and robust accumulation process of evidence to the BG function.

However, we also note that, in contrast to our previous studies (Frank *et al.*, 2007, 2009b; Doll *et al.*, 2011), we did not observe an effect of SNPs impacting striatal dopaminergic function (DARPP-32 and DRD2) on RL parameters. However, the present experiment was designed to specifically assess the effects of WM on learning much more than the effects of incremental BG learning. In particular, feedback was binary and deterministic. In contrast, previous experiments reporting effects of these SNPs have all required subjects to discriminate between subtly different reinforcement probabilities of positive and negative outcomes. A direction for future research would be to test the hybrid WM and RL model in an experiment allowing equal differentiation on both systems. Given the selective presence of GPR6 in the indirect pathway (Lobo *et al.*, 2007), and computational models of this pathway, we can predict that this SNP would be specifically related to individual differences in learning from negative prediction errors (Frank, 2005; Frank & Fossella, 2010).

Limitations

The proposed WM part in the RL + WM model proposed remains very simple. Although we hypothesize a limited capacity, we do not attempt to identify which specific events are stored in memory and which are not, but simply account for the probability of any one of them being stored, given capacity limits and memory decay. This probability is modeled in the simplest way, with an equal probability of each stimulus being stored in memory. Although this allowed us to capture most WM effects in a simple model, it is an approximation that deserves more attention in future research. For example, this probability might dynamically change over time, with the possibility of an event being replaced in memory by other intervening stimuli. Indeed, it is possible that such updating effects are in part responsible for the delay effects within the WM module, currently represented as a simple passive decay. More work is thus needed to distinguish this possibility.

Although our results and discussion may imply that the BG is selectively involved in incremental motor action value learning, in other work we have emphasized an analogous role of other BG circuits in the learning of when and when not to gate information into WM (Frank *et al.*, 2001; O'Reilly & Frank, 2006; Frank & Badre, 2011; Collins & Frank, unpublished data; see also Todd *et al.*, 2008; Gruber *et al.*, 2006). This role of the BG in updating WM is also supported by empirical studies with patients, pharmacological manipulations, and neuroimaging (Cools *et al.*, 2007; Moustafa *et al.*, 2008; Baier *et al.*, 2010). As in most RL tasks, the current task does not impose demands on learning which stimulus information is relevant to update and which is not. We anticipate that neurogenetic studies will reveal that learning when and when not to gate information into WM will load on BG genes (Frank & Fossella, 2010). Indeed, some evidence indicates that improvements in performance due to cognitive training on WM updating tasks are related to striatal activation (Dahlin *et al.*, 2008).

Conclusion

In this study, we have illustrated the necessity to account for higher cognitive contributions to instrumental learning paradigms. Indeed, we showed that simple RL models of these tasks failed to account for the observed effects of memory load and time delay on performance. Moreover, incorporating these effects into the model was crucial for properly assigning behavioral variance to its proper causes when relating behavior to neurogenetic correlates. This new model, which dynamically balanced both PFC–WM and BG–RL aspects in learning, allowed us to show that, although learning was affected by both prefrontal function and BG function markers, the former is attributable to variance in WM capacity, whereas the latter is attributable to the overall incremental learning speed.

Abbreviations

AIC, Akaike information criterion; BG, basal ganglia; COMT, catechol-O-methyl transferase; RL, reinforcement learning; SNP, single nucleotide polymorphism; WM, working memory.

References

Baier, B., Karnath, H.-O., Dieterich, M., Birklein, F., Heinze, C. & Muller, N.G. (2010) Keeping memory clear and stable – the contribution of human basal ganglia and prefrontal cortex to working memory. *J. Neurosci.*, **30**, 9788–9792.

Bayer, H.M. & Glimcher, P.W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, **47**, 129–141.

Bays, P.M. & Husain, M. (2008) Dynamic shifts of limited working memory resources in human vision. *Science*, **321**, 851–854. doi: 10.1126/science.1158023.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E. & Rushworth, M.F.S. (2007) Learning the value of information in an uncertain world. *Nat. Neurosci.*, **10**, 1214–1221.

Blasi, G., Mattay, V.S., Bertolino, A., Elvevåg, B., Callicott, J.H., Das, S., Kolachana, B.S., Egan, M.F., Goldberg, T.E. & Weinberger, D.R. (2005) Effect of catechol-O-methyltransferase val158met genotype on attentional control. *J. Neurosci.*, **25**, 5038–5045.

Botvinick, M., Niv, Y. & Barto, A. (2009) Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. *Cognition*, **113**, 262–280.

Brovelli, A., Laksiri, N., Nazarian, B., Meunier, M. & Boussaoud, D. (2008) Understanding the neural computations of arbitrary visuomotor learning through fMRI and associative learning theory. *Cereb. Cortex*, **18**, 1485–1495.

Bruder, G.E., Keilp, J.G., Xu, H., Shikhman, M., Schori, E., Gorman, J.M. & Gilliam, T.C. (2005) Catechol-O-methyltransferase (COMT) genotypes and working memory: associations with differing cognitive operations. *Biol. Psychiatry*, **58**, 901–907.

Cavanagh, J.F., Frank, M.J., Klein, T.J. & Allen, J.J.B. (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*, **49**, 3198–3209.

Cools, R., Sheridan, M., Jacobs, E. & D'Esposito, M. (2007) Impulsive personality predicts dopamine-dependent changes in frontostriatal activity during component processes of working memory. *J. Neurosci.*, **27**, 5506–5514.

Corrado, G. & Doya, K. (2007) Understanding neural coding through the model-based analysis of decision making. *J. Neurosci.*, **27**, 8178–8180.

Cowan, N. (2010) The magical mystery four: how is working memory capacity limited, and why? *Curr. Dir. Psychol. Sci.*, **19**, 51–57.

Dahlin, E., Neely, A.S., Larsson, A., Bäckman, L. & Nyberg, L. (2008) Transfer of learning after updating training mediated by the striatum. *Science*, **320**, 1510–1512.

Daw, N.D. (2011) Trial-by-trial data analysis using computational models. In Phelps, E.A., Robbins, T.W. & Delgado, M. (Eds), *Affect, Learning and Decision Making, Attention and Performance XXIII*. Oxford University Press, New York, pp. 3–38.

Daw, N.D. & Doya, K. (2006) The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.*, **16**, 199–204.

Daw, N.D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorso-lateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. & Dolan, R.J. (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron*, **69**, 1204–1215.

Doll, B.B., Hutchison, K.E. & Frank, M.J. (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.*, **31**, 6188–6198.

Doya, K. (2002) Metalearning and neuromodulation. *Neural Netw.*, **15**, 495–506.

Durstewitz, D. & Seamans, J.K. (2008) The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. *Biol. Psychiatry*, **64**, 739–749.

Durstewitz, D., Vittoz, N.M., Floresco, S.B. & Seamans, J.K. (2010) Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron*, **66**, 438–448.

Egan, M.F., Goldberg, T.E., Kolachana, B.S., Callicott, J.H., Mazzanti, C.M., Straub, R.E., Goldman, D. & Weinberger, D.R. (2001) Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc. Natl. Acad. Sci. USA*, **98**, 6917–6922.

Ernst, C., Sequeira, A., Klempan, T., Ernst, N., ffrench-Mullen, J. & Turecki, G. (2007) Confirmation of region-specific patterns of gene expression in the human brain. *Neurogenetics*, **8**, 219–224.

Evans, J.S.B.T. (2003) In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.*, **7**, 454–459.

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

Frank, M.J. & Badre, D. (2011) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb. Cortex*, doi:10.1093/cercor/bhr114.

Frank, M.J. & Fossella, J.A. (2010) Neurogenetics and pharmacology of learning, motivation, and cognition. *Neuropsychopharmacology*, **36**, 133–152.

Frank, M.J., Loughry, B. & O'Reilly, R.C. (2001) Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cogn. Affect. Behav. Neurosci.*, **1**, 137–160.

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

Frank, M., Doll, B., Oas-Terpstra, J. & Moreno, F. (2009a) The neurogenetics of exploration and exploitation Supplementary Material. *Nature Neuroscience*, **12**, 1062.

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

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

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

Goldberg, T.E. & Weinberger, D.R. (2004) Genes and the parsing of cognitive processes. *Trends Cogn. Sci.*, **8**, 325–335.

Green, A., Munafò, M., DeYoung, C., Fossella, J., Fan, J. & Gray, J. (2008) using genetic data in cognitive neuroscience: from growing pains to genuine insights. *Nat. Rev. Neurosci.*, **9**, 710–720.

Gruber, A.J., Dayan, P., Gutkin, B.S. & Solla, S.A. (2006) Dopamine modulation in the basal ganglia locks the gate to working memory. *J. Comput. Neurosci.*, **20**, 153–166.

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

Huotari, M., Gogos, J.A., Karayiorgou, M., Koponen, O., Forsberg, M., Raasmaja, A., Hyttinen, J. & Männistö, P.T. (2002) Brain catecholamine metabolism in catechol-O-methyltransferase (COMT)-deficient mice. *Eur. J. Neurosci.*, **15**, 246–256.

Imamizu, H., Sugimoto, N., Osu, R., Tsutsui, K., Sugiyama, K., Wada, Y. & Kawato, M. (2007) Explicit contextual information selectively contributes to predictive switching of internal models. *Exp. Brain Res.*, **181**, 395–408.

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

Kahnt, T., Park, S.Q., Cohen, M.X., Beck, A., Heinz, A. & Wrase, J. (2009) Dorsal striatal-midbrain connectivity in humans predicts how reinforcements are used to guide decisions. *J. Cogn. Neurosci.*, **21**, 1332–1345.

Lau, B. & Glimcher, P.W. (2007) Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.*, **27**, 14502–14514. doi: 10.1523/JNEUROSCI.3060-07.2007.

Lengyel, M. & Dayan, P. (2007) Hippocampal contributions to control: the third way. *NIPS*, **20**, 889–896.

Lobo, M.K., Cui, Y., Ostlund, S.B., Balleine, B.W. & Yang, X.W. (2007) Genetic control of instrumental conditioning by striatopallidal neuron-specific S1P receptor Gpr6. *Nat. Neurosci.*, **10**, 1395–1397.

Matsumoto, M., Weickert, C.S., Akil, M., Lipska, B.K., Hyde, T.M., Herman, M.M., Kleinman, J.E. & Weinberger, D.R. (2003) Catechol O-methyltransferase mRNA expression in human and rat brain: evidence for a role in cortical neuronal function. *Neuroscience*, **116**, 127–137.

Mier, D., Kirsch, P. & Meyer-Lindenberg, A. (2010) Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis. *Mol. Psychiatry*, **15**, 918–927.

Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947.

Moustafa, A.A., Sherman, S.J. & Frank, M.J. (2008) A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism. *Neuropsychologia*, **46**, 3144–3156.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R.J. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, **304**, 452–454.

O'Doherty, J.P., Hampton, A. & Kim, H. (2007) Model-based fMRI and its application to reward learning and decision making. *Ann. NY Acad. Sci.*, **1104**, 35–53.

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

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J. & Frith, C.D. (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, **442**, 1042–1045.

Ribas-Fernandes, J.J.F., Solway, A., Diuk, C., McGuire, J.T., Barto, A.G., Niv, Y. & Botvinick, M.M. (2011) A neural signature of hierarchical reinforcement learning. *Neuron*, **71**, 370–379.

Roth, R.B., Hevezi, P., Lee, J., Willhite, D., Lechner, S.M., Foster, A.C. & Zlotnik, A. (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, **7**, 67–80.

Samejima, K., Ueda, Y., Doya, K. & Kimura, M. (2005) Representation of action-specific reward values in the striatum. *Science*, **310**, 1337–1340.

Schönberg, T., Daw, N.D., Joel, D. & O'Doherty, J.P. (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.*, **27**, 12860–12867.

Schultz, W. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.

Sesack, S.R., Hawrylak, V.A., Matus, C., Guido, M.A. & Levey, A.I. (1998) Dopamine axon varicosities in the prelimbic division of the rat prefrontal cortex exhibit sparse immunoreactivity for the dopamine transporter. *J. Neurosci.*, **18**, 2697–2708.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R., Friston, K. & Frackowiak, R. (2004) Temporal difference models describe higher-order learning in humans. *Nature*, **429**, 664–667.

Slifstein, M., Kolachana, B., Simpson, E.H., Tabares, P., Cheng, B., Duvall, M., Frankle, W.G., Weinberger, D.R., Laruelle, M. & Abi-Dargham, A. (2008) COMT genotype predicts cortical-limbic D1 receptor availability measured with [11C]NNC112 and PET. *Mol. Psychiatry*, **13**, 821–827.

Sloman, S.A. & Hagmayer, Y. (2006) The causal psycho-logic of choice. *Trends Cogn. Sci.*, **10**, 407–412.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J. & Friston, K.J. (2009) Bayesian model selection for group studies. *NeuroImage*, **46**, 1004–1017.

Sutton, R. & Barto, A. (1998) *Reinforcement Learning*, Vol. **9**. MIT Press, Cambridge, Massachussetts.

Tan, H.-Y., Chen, Q., Goldberg, T.E., Mattay, V.S., Meyer-Lindenberg, A., Weinberger, D.R. & Callicott, J.H. (2007) Catechol-O-methyltransferase Val158Met modulation of prefrontal-parietal-striatal brain systems during arithmetic and temporal transformations in working memory. *J. Neurosci.*, **27**, 13393–13401.

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y. & Yamawaki, S. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat. Neurosci.*, **7**, 887–893.

Tanaka, S.C., Samejima, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S. & Doya, K. (2006) Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Netw.*, **19**, 1233–1241.

Todd, M.T., Niv, Y. & Cohen, J.D. (2008) Learning to use working memory in partially observable environments through dopaminergic reinforcement. in Koller, D. (Ed.), Twenty-First Annual Conference on Neural Information Processing Systems (NIPS) 2008. Vancouver: Canada.

Tunbridge, E.M., Bannerman, D.M., Sharp, T. & Harrison, P.J. (2004) Catechol-o-methyltransferase inhibition improves set-shifting performance and elevates stimulated dopamine release in the rat pre-frontal cortex. *J. Neurosci.*, **24**, 5331–5335.

Zhang, W. & Luck, S.J. (2008) Discrete fixed-resolution representations in visual working memory. *Nature*, **453**, 233–235.

Zhang, W. & Luck, S.J. (2011) The number and quality of representations in working memory. *Psychol. Sci.*, **22**, 1434–1441. doi: 10.1177/0956797611417006.