Cerebral Cortex March 2012;22:509–526
doi:10.1093/cercor/bhr114
Advance Access publication June 21, 2011

Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis

Michael J. Frank and David Badre

Department of Cognitive, Linguistic Sciences and Psychological Sciences, Brown Institute for Brain Science, Brown University, Providence RI 02912-1978, USA

Address correspondence to email: michael_frank@brown.edu.

Growing evidence suggests that the prefrontal cortex (PFC) is organized hierarchically, with more anterior regions having increasingly abstract representations. How does this organization support hierarchical cognitive control and the rapid discovery of abstract action rules? We present computational models at different levels of description. A neural circuit model simulates interacting corticostriatal circuits organized hierarchically. In each circuit, the basal ganglia gate frontal actions, with some striatal units gating the inputs to PFC and others gating the outputs to influence response selection. Learning at all of these levels is accomplished via dopaminergic reward prediction error signals in each corticostriatal circuit. This functionality allows the system to exhibit conditional if–then hypothesis testing and to learn rapidly in environments with hierarchical structure. We also develop a hybrid Bayesian-reinforcement learning mixture of experts (MoE) model, which can estimate the most likely hypothesis state of individual participants based on their observed sequence of choices and rewards. This model yields accurate probabilistic estimates about which hypotheses are attended by manipulating attentional states in the generative neural model and recovering them with the MoE model. This 2-pronged modeling approach leads to multiple quantitative predictions that are tested with functional magnetic resonance imaging in the companion paper.

Keywords: basal ganglia, computational model, hierarchical reinforcement learning, prefrontal cortex

Introduction

Flexible behavior requires cognitive control, or the ability to select mappings between states and actions based on internally maintained representations of context, goals, and anticipated outcomes (Miller and Cohen 2001; O'Reilly and Frank 2006; Balleine and O'Doherty 2010). Often, however, control over action requires simultaneously maintaining features of the context that are relevant for action selection at different levels of abstraction and over varying timescales (Koechlin and Summerfield 2007; Badre 2008; Botvinick 2008). Growing evidence has suggested that the frontal cortex is functionally organized to support this type of hierarchical control, such that progressively rostral portions of frontal cortex govern control at higher levels of abstraction (Koechlin et al. 2000, 2003; Christoff et al. 2003, 2009; Koechlin and Hyafil 2007; Badre and D'Esposito 2007; Badre et al. 2009; Kouneiher et al. 2009).

Recent evidence has indicated that when humans are confronted with a new rule learning problem, this rostrocaudal division of labor in frontal cortex supports the search for relationships between context and action at multiple levels of abstraction simultaneously (Badre et al. 2010). Human participants were scanned with functional magnetic resonance imaging (fMRI) while performing a reinforcement learning task in which they learned 18 mappings between the conjunction of 3 features of a presented stimulus (shape, orientation, and color) and one of 3 finger responses on a key pad. Critically, each participant learned 2 such sets of 18 rules. For one of these sets, an abstract rule was available that would permit generalization across multiple individual mappings of stimuli and responses. The results from this experiment demonstrated that 1) participants were capable of rapidly discovering and applying the abstract rule when it was available. 2) fMRI activation was evident in both dorsal premotor cortex (PMd) and more rostral premotor cortex (prePMd) early in learning but declined in the prePMd by the end of learning when no abstract rule was available. 3) Individual differences in the activation early in learning in prePMd, but not in PMd, were correlated with participants' likelihood of discovering an abstract rule when one was available. And, 4) striatum (caudate and putamen) showed greater activation by the end of learning during the session when abstract rules were available. However, the network dynamics of frontostriatal interactions, as estimated by functional connectivity, did not differ based on the presence or absence of a higher order rule. Hence, these results suggest that from the outset of learning the search for relationships between context and action may occur at multiple levels of abstraction simultaneously and that this process differentially relies on systematically more rostral portions of frontal cortex for the discovery of more abstract relationships.

A key question motivated by this work, concerns what neural mechanisms support this higher order rule discovery? Here, we first consider whether an existing neural model of corticostriatal circuits in reinforcement learning and working memory (Frank 2005; O'Reilly and Frank 2006) provides a plausible set of mechanisms, when they are modified to accommodate hierarchical structure. In this model, the striatum modulates the selection of frontal cortical actions, including motor actions and working memory updating. It does so by gating the inputs to be maintained in frontal cortex (input gating) and gating which of these maintained representations has an influence on action selection (output gating). The learning of which actions are adaptive given a particular state is accomplished via dopaminergic reinforcement learning mechanisms. The basic properties of this model (and of related models) have been extensively investigated elsewhere (Frank 2005; O'Reilly and Frank 2006; Hazy et al. 2007; Reynolds and O'Reilly 2009). Here, in order to support learning of hierarchical structure within the context of the Badre et al. (2010) task, we modified the model such that anterior regions of prefrontal cortex (PFC) contextualize striatal gating of more posterior frontal regions. More specifically, as with previous models (O'Reilly & Frank 2006; Frank and Claus 2006; Reynolds and

O'Reilly 2009), working memory representations maintained in the "posterior" PFC layer of the model constrain motor output decisions. Unique to the present model, working memory representations maintained in additional "anterior" PFC layers constrain which of the working memory representations maintained in posterior PFC should be "output-gated" to influence attention and ultimately response selection. We show that this model can improve performance in tasks requiring acquisition of a hierarchical structure relative to a model that has no such functionality and that it does so by learning an abstract gating policy.

We then derive a more abstract Bayesian-reinforcement learning (RL) mixture of experts (MoE) model intended to correspond to key computational features of the neural model. We use this model to estimate latent states, that is, hypotheses about the relationships between context and action that are most likely being tested, in individual human learners given their trial-by-trial sequences of choices and rewards. We also investigate the relationship between these probabilistic estimates and the mechanisms that govern hypothesis testing and learning in the generative neural circuit model, treating the latter model as a participant and fitting its choices with the MoE. We find that the fits to the neural model are similar to those to human participants, despite the fact that the neural model stochastically gates a subset of features in individual trials (i.e., it operates according to one hypothesis or another), whereas the MoE model assumes a probabilistic mixture of hypotheses on all trials. We further report that the degree of learning, and the MoE estimate of attention to hierarchical structure, is correlated with a measure of gating policy abstraction in the learned weights of the neural model. The combined set of analyses provides the theoretical basis for the multiple model predictions tested with functional imaging analysis in the companion paper.

Corticostriatal Mechanisms of Action Selection and Hierarchical Reinforcement Learning

Several corticostriatal models take as their starting point the general notion that the basal ganglia (BG) act as a "gate" to facilitate particular action plans in frontal cortex while suppressing other less adaptive plans (e.g., Mink 1996; Gurney et al. 2001a; Brown et al. 2004; Frank 2005). Motivation for such an architecture comes partly from evidence for parallel frontal corticostriatal loops (Alexander et al. 1986). In the motor domain, based on sensory information, the premotor cortex first selects candidate motor responses and then the motor BG selectively amplifies representations of one these candidates (Frank 2005). In the cognitive domain, the BG select which candidate stimuli to selectively update and subsequently maintain in PFC (Frank et al. 2001; Houk 2005; O'Reilly and Frank 2006; Gruber et al. 2006). Computational trade-offs indicate that it is adaptive to have separate systems implement gating and maintenance (Hochreiter and Schmidhuber 1997) and that the functional BG circuitry is well suited to selectively gate particular working memory representations while allowing others to continue to be maintained. In turn, maintained PFC representations project to the motor BG such that motor response selection is sensitive to the combination of both input and PFC states (Frank et al. 2001).

In both motor and cognitive domains, the selection of which actions to facilitate, and which to suppress, is learned via a common dopaminergic reward prediction error signal that modulates activity in "Go" and "NoGo" striatal neuronal populations expressing D1 and D2 dopamine receptors, respectively (Frank 2005; O'Reilly and Frank 2006; Shen et al. 2008). Multiple lines of evidence support this role of the BG and dopamine in selection and learning across both motor and cognitive domains, including effects of BG and PFC brain damage (Baier et al. 2010; Voytek and Knight 2010), correlates in functional imaging (McNab and Klingberg 2008; Cools et al. 2008), pharmacological manipulations in patients with Parkinson's disease (Frank et al. 2004; Cools et al. 2001, 2006; Moustafa et al. 2008; Dagher and Robbins 2009; Palminteri et al. 2009), the combination of pharmacology and/or Parkinson's patients with neuroimaging (Pessiglione et al. 2006; Siessmeier et al. 2006; Cools, Lewis, et al. 2007; Cools, Sheridan, et al. 2007; Cools et al. 2009; Schonberg et al. 2010; Voon et al. 2010), and finally, pharmacology and genetics of striatal dopamine function in young healthy individuals (Frank and O'Reilly 2006; Stollstorff et al. 2010; Frank and Fossella 2011).

In cognitive tasks, it may be necessary to update and maintain multiple task-relevant items in working memory. In the corticostriatal network models, stimuli are selectively "input-gated" into segregated PFC memory "stripes" (structures of interconnected neurons that are isolated from other adjacent stripes; Pucak et al. 1996; Frank et al. 2001). In this manner, motor response selection can then be contextualized by multiple potential working memory representations. However, in some scenarios only a limited subset of currently maintained PFC representations may be relevant for influencing the current motor decision (e.g., in tasks with multiple goals and subgoals, only some maintained representations are relevant for processing during intermediate stages). In such cases, the system can benefit from an additional "output-gating" mechanism, whereby an independent set of striatal units select which, among multiple currently maintained PFC memory representations, should influence the current motor response selection (Fig. 2) (Hochreiter and Schmidhuber 1997; Brown et al. 2004; Hazy et al. 2007). (Output-gating relies on the very same mechanisms as those supporting gating of motor actions, whereby striatal units disinhibit thalamic units that interact bidirectionally with deep (layer 5) PFC units which in turn send descending projections to subcortical regions, (e.g., the motor BG) and to posterior cortex (to bias attention). In contrast, striatal input gating units modulate the maintenance of superficial PFC representations (layers 2/3).)

Note that the decision of which of the current PFC memory representations should be output-gated may be contingent on the identity of some other PFC representation (Hochreiter and Schmidhuber 1997). (In the language of Hochreiter and Schmidhuber (1997), "an output gate may use inputs from other memory cells to decide whether to access certain information in its memory cell" (p. 1744).) We posit that this higher level contextual information would be represented in more anterior PFC regions (Fuster 1997; Botvinick 2007; Badre 2008; Reynolds and O'Reilly 2009). Just as a PFC working memory representations influence the output gating of BG motor responses, more anterior PFC representations can influence the BG output gating units of the posterior PFC representations (Figure 2, right). Such a hierarchical organization of corticostriatal circuits may facilitate adaptive cognitive and motor behavior over multiple timescales and levels of cognitive complexity (see Discussion).

A critical question is how to determine (without prior instruction) which motor responses to select, which stimulus

dimensions are relevant for directly constraining motor responses, and which dimensions should be considered "higher order" contexts? A large body of literature suggests that the corticostriatal decisions of whether or not to gate particular actions is contingent on the reward-predictive properties of these actions. In particular, midbrain dopamine neurons signal when rewards are better or worse than expected (Montague et al. 1997; Schultz et al. 1997), and these phasic signals drive learning in Go and NoGo gating units via dopaminergic modulation of synaptic plasticity (Reynolds and Wickens 2002; Frank 2005; Shen et al. 2008). For reviews of the evidence for these mechanisms across species and methodologies, see Frank and Fossella (2011; Surmeier et al. 2010).

In the neural circuit model used here, adapted from O'Reilly and Frank (2006), relatively anterior corticostriatal circuits learn via dopaminergic signals which stimuli are predictive of reward when considered as "contextual." When gated into anterior PFC, these stimuli serve to constrain the decision of which of the other stimulus dimensions to attend in more posterior circuits: anterior PFC neurons project to striatal units that perform output-gating of posterior PFC. Supporting this functionality, recent anatomical evidence shows a substantial degree of convergence between corticostriatal circuits (Haber 2004; Calzavara et al. 2007; Draganski et al. 2008), and in particular, evidence for a rostrocaudal organization from premotor/PFC to corresponding regions of striatum (Inase et al. 1999; Lehericy, Ducros, Krainik, et al. 2004; Lehericy, Ducros, Vande Moortele, et al. 2004; Postuma and Dagher 2006; Draganski et al. 2008). Moreover, the general notion that multiple corticostriatal loops are involved in different sorts of belief states and action values is largely consistent with that proposed in a recent review (Samejima and Doya 2007). In particular, those authors proposed that the lateral prefrontal-anterior striatal circuit is involved in context-based value estimation.

Materials and Methods

We implemented computational models at 2 levels of description. The first builds on existing neural models of corticostriatal circuits in reinforcement learning and working memory and extends this framework to accommodate hierarchical structure. This model attempts to provide a mechanistic understanding of how such circuitry is recruited to facilitate the discovery of hierarchical structure in the environment. The second model develops a higher level abstract analysis of the processes engaged during the learning of a specific hierarchical task but where some of the model assumptions are motivated by the core computations in the neural model. This model attempts to provide an account of individual learners and its free parameters are varied to maximize fit to trial-by-trial sequences of choices, allowing us to infer whether learners are most likely to be testing hierarchical (or other) hypotheses about task structure. In the companion paper, we utilize the higher level model to derive regressors to examine functional neuroimaging data and interpret these data in the context of the mechanisms specified by the neural model.

We sought to apply both models to simulate behavior and neural dynamics from a recently reported hierarchical reinforcement learning task (Badre et al. 2010). During fMRI scanning, participants were required to learn 2 sets of rules, in separate epochs, that linked each of 18 different stimulus conjunctions uniquely and deterministically to one of 3 button press responses (Fig. 1). For each rule set, an individual stimulus conjunction consisted of one of 3 shapes, at one of 3 orientations, inside a box that was one of 2 colors, for a total of 18 unique stimuli (3 shapes 3 orientations × 2 colors). Participants were instructed to learn the correct response for each stimulus based on

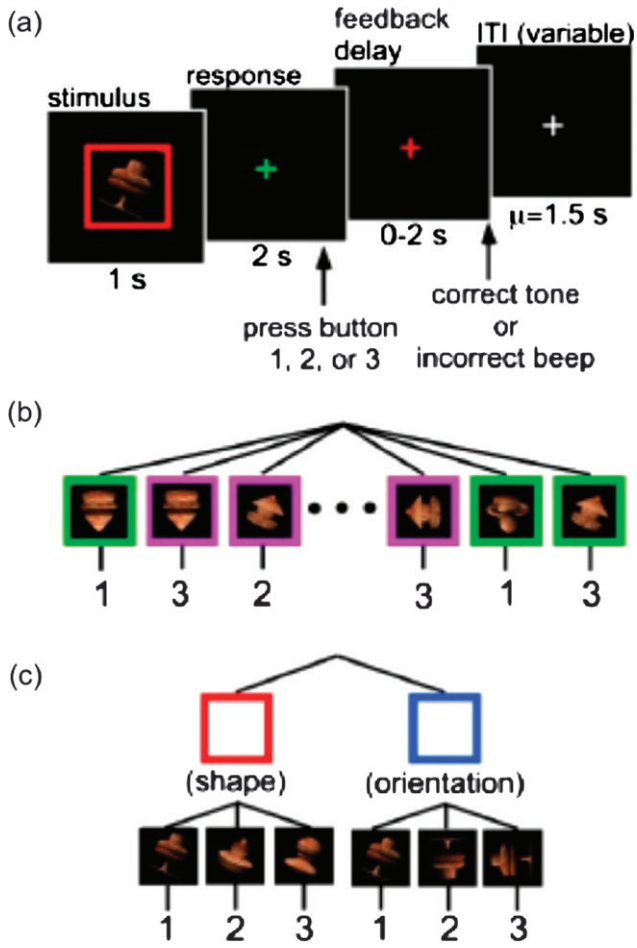

Figure 1. Badre et al. (2010) hierarchical reinforcement learning task. A schematic depiction of trial events along with example stimulus-to-response mappings for hierarchical and flat rule sets. (*a*) Trials began with presentation of a stimulus followed by a green fixation cross. Participants could respond with a button press at any time while the stimulus or green fixation cross was present. After a variable delay following the response, participants received auditory feedback indicating whether the response they had chosen was correct given the presented stimulus. Trials were separated by a variable null interval. (*b*) Example stimulus-to-response mappings for the Flat set. The arrangement of mappings for the Flat set was such that no higher order relationship was present; thus, each rule had to be learned individually. (*c*) Example stimulus-to-response mappings for the Hierarchical set. Response mappings in this example are grouped such that in the presence of a red square, only shape determines the response, while in the presence of a blue square, only orientation determines the response.

auditory reinforcement feedback (they were also paid in proportion to the number of correct responses). For one of the 2 rule sets (Flat set), each of the 18 rules had to be learned individually as one-to-one mappings between a conjunction of color, shape, and orientation and a response. In the other set (Hierarchical set), stimulus display parameters and instructions were identical to the Flat set. And, indeed, the Hierarchical set could also be learned as 18 first-order rules. However, the arrangement of response mappings was such that a second-order relationship could be learned instead. In the context of one colored box, only the shape dimension was relevant to the response, with each of the 3 unique shapes mapping to one of the 3 button responses regardless of orientation. Conversely, in the context of the other colored box, only the orientation dimension was relevant to the response. Thus, the Hierarchical rule set permitted learning of a more abstract conditional rule that specified how one dimension (color) determined which of the other dimensions (shape or orientation) would provide a context for selecting a response. Again, all instructions, stimulus presentation parameters, and between-subject stimulus orderings were identical between the 2 rule sets. The Flat and

Hierarchical rule sets only differed in that the organization of mappings in the Hierarchical set permitted learning of a more abstract rule. Hence, these 2 sets contrast a learning context in which abstract rules can be discovered with an analogous context in which no such rules can be learned. Thus, this design provides a means of studying the neural mechanisms of abstract rule learning.

Corticostriatal Neural Circuit Model

The focus of this paper is not on the neural model itself, which has been investigated elsewhere (Frank 2005; O'Reilly and Frank 2006). For the sake of brevity and focus on the primary algorithmic model description, we report all neural model methods, equations for individual neuron activation dynamics, reinforcement learning rules in the BG, and prefrontal working memory mechanisms in the Supplementary Material. Here, we provide a high level summary of its application in this context.

A schematic of the neural model structure is presented in Figure 2. The 3 stimulus dimensions were represented as simple localist units (3 units each for shape and orientation and 2 units for color) and provided as input to the network. We included 2 separate frontal layers (corresponding to PMd and prePMd). Each layer had 3 "stripes" which had the capacity to represent each of the stimulus dimensions. This implementation is consistent with the notion that each part of a multifeatural object is represented in a separate slot in visual working memory (Xu 2002; Sakai and Inui 2002). This structure was duplicated in maintenance layers (which maintain information over delays) and output layers (which convey output-gated information to the response selection network). The central addition to the model used here is that it includes multiple circuits, whereby the more anterior (prePMd) frontal layer also provides contextual input to the striatal output-gating layers of the more posterior (PMd) region. Thus, the decision of whether or not to attend to a particular stimulus dimension in PMd, and ultimately whether that dimension is used for motor response selection, is contextualized by information maintained in prePMd. Dopaminergic signals convey reward prediction errors that modulate synaptic plasticity and hence learning in all striatal units, such that increases in dopamine promote Go learning and decreases in dopamine promote NoGo learning. These mechanisms allow networks to discover which stimulus dimensions are predictive of reward if gated into PMd or prePMd, while simultaneously learning the specific stimulus–response mappings. In particular, networks should learn that in the hierarchical condition, color should be gated into prePMD, and the striatal gating units in the PMd circuit should learn to contextualize which of the other dimensions to output-gate depending on the prePMD representation.

For each frontal stripe, the corresponding striatal gating layers consisted of 28 distributed units (14 Go and 14 NoGo) which learn the probability of obtaining a reward if the stimulus in question (e.g., a particular shape) is gated into, or out of, its respective working memory stripe. In each module, an SNr/Thal (substantia nigra/thalamus) unit implements a gating signal and is activated when relatively more striatal Go than NoGo units are active (subject to inhibitory competition from other SNr/Thal units that modulate gating of neighboring stripes (O'Reilly and Frank 2006). Thus, the SNr/Thal units summarize the contributions of multiple interacting layers that implement gating among the substantia nigra, globus pallidus, subthalamic nucleus, and thalamus as simulated in more detailed networks of a single BG circuit (Frank 2006), in these larger-scale networks we abstract away from these details. For input-gating circuits, SNr/Thal activation induces maintenance of activation states in the corresponding frontal maintenance layer (Frank et al. 2001; O'Reilly and Frank 2006). For output-gating circuits, the SNr/Thal activation results in information flow from the frontal maintenance layer to the frontal output layer. This output layer projects to the decision circuit, such that only output-gated representations influence response selection (see Fig. 2).

To simulate the Badre et al. task, the model was trained with the same stimulus–response contingencies administered to the subjects, in pseudorandom order for 400 trials in each condition (hierarchical and flat). Each trial consisted of stimulus presentation, during which stimuli could be gated into corresponding PFC areas, followed by another phase in which all input stimuli were removed and the network had to rely on maintained PFC representations in order to respond. (This working memory aspect was included to capture a design feature employed in fMRI in which stimuli were presented for a brief period and were then removed from the display before participants responded in order to equate visual presentation time independent of response time.) The frontal stripes for each of the stimulus dimensions could independently maintain representations of these stimulus dimensions in PMd_{Maint}, subject to gating signals from the BG. Initially, a "Go bias" encourages exploratory updating (and subsequent maintenance) due to novelty; these gating signals are then reinforced to the extent that the frontal representations come to be predictive of reward (O'Reilly and Frank 2006). However, not all maintained PMd representations influence decision in the response circuitry – only those that are also represented in PMD_{Out} due to output gating signals. Thus, in a given trial, shape and orientation of the current stimulus may be represented in PMD_{Maint} but depending on output gating, for example, only the shape will be represented in PMD_{Out} and thereby influence the motor decisions. To facilitate the discovery of hierarchical structure, we included projections from the more anterior prePMd layer to the striatal output gating units of PMd. Dopaminergic reinforcement signals operate at all these levels, supporting gating of information that will be most useful for constraining response selection. Thus, the input-gating units to prePMd should learn to gate in the color dimension to prePMd, which can then act to contextualize the output-gating decision of PMd and ultimately response selection. Critically, this scheme prevents the motor response selection network from having to learn multiple conjunctive stimulus–response associations and prevents interference between similar stimuli with opposing motor actions. For example, shape stimuli are only considered for response selection for one of the contexts and do not interfere with response associations when the other context is present. For further details on these simulations, please see the Supplementary Material.

Bayesian-RL MoE Model

We now develop a more abstract computational-level account of the learning process, intended to capture key computational features of the neural model but suitable for analysis of individual learners. Whereas the neural model focuses on a plausible implementation in interacting networks of neurons across multiple brain areas and can produce qualitative fits to the data, it is not appropriate for fitting trial-to-trial sequence of choices in any given subject and it is not clear which parameters would be allowed to vary freely (for similar arguments and methods comparing neural and abstract RL and Bayesian learning models, see Doll et al. (2009)). We aimed to develop a model which could estimate the hypotheses that a subject might be testing in a given trial—that is, their "latent states"—which we could estimate only by observing the sequence of stimuli, responses, and rewards that the subject experienced and then maximizing the likelihood of their observed choices under the model given these observations. These simulations involved few free parameters, which correspond to intuitive ways in which participants might differ, for example, the prior likelihood of attending to shape versus orientation versus color.

We modeled individual learners using an MoE architecture (see Fig. 3). We adopt a hybrid Bayesian-RL formulation here (see below section on model fitting for discussion of motivation for use of a Bayesian learning rule in updating attentional weights to account for prior biases). Each expert focuses on a particular stimulus dimension or combination of dimensions and learns the probability of obtaining a reward for each motor response given the features present in their domain of expertise. For example, the orientation expert would learn $P(\text{Rew}|\text{Response}, \text{Orient})$, and so forth for other experts.

As the outcomes were binomial, for each expert, we modeled participants' belief about reward probability for each of the 3 responses as a beta distribution $\text{Beta}(\alpha, \beta)$ (see Supplementary Material) which was updated as a function of experience via Bayes' Rule, that is, the posterior distribution about reward probability for each response was updated as a function of its prior distribution and the likelihood that the particular sequence of rewards would be observed given this prior. For the orientation expert, this update is given by:

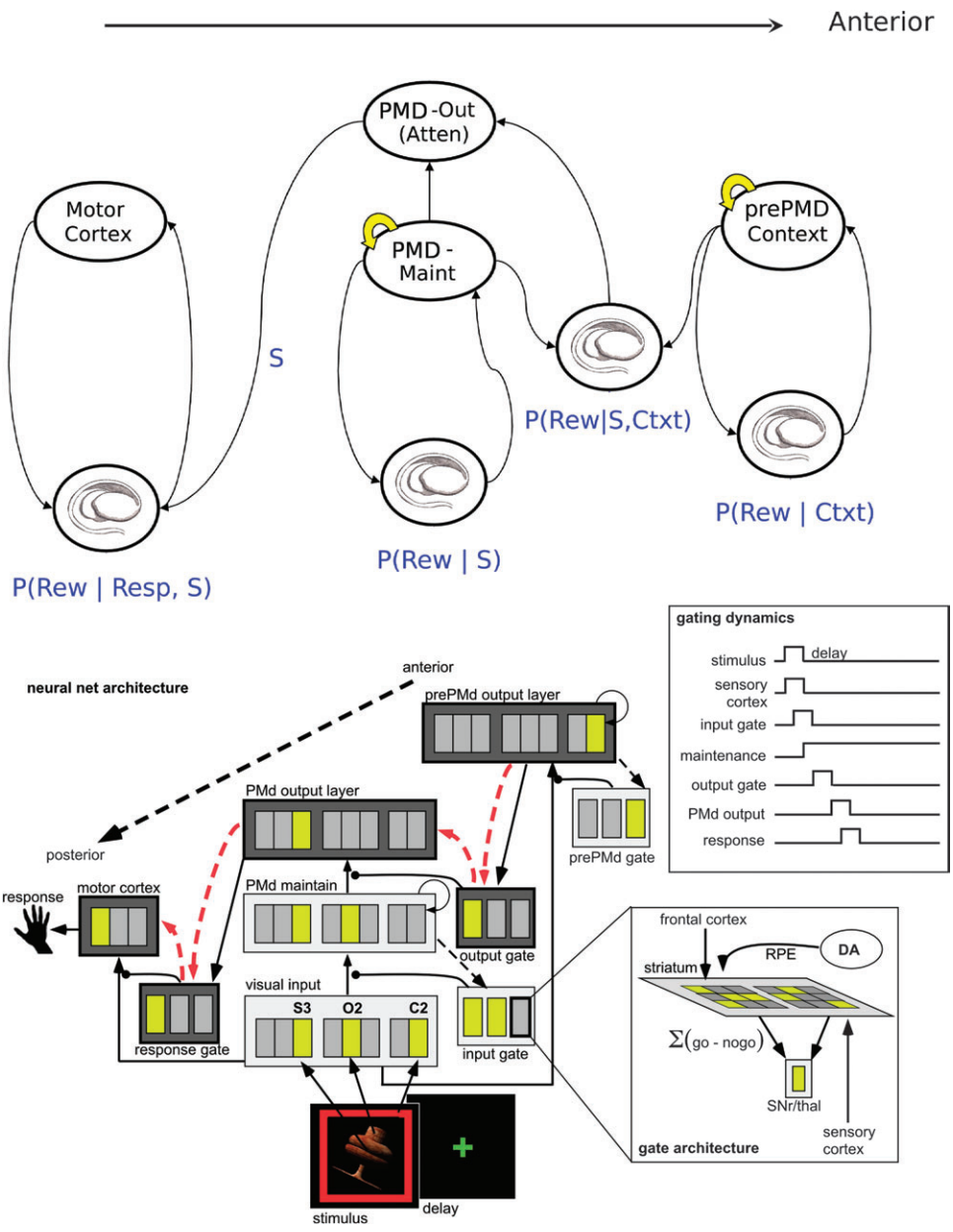

Figure 2. Schematic of hierarchical corticostriatal circuit. In the standard response selection circuit, motor areas of the striatum interact with motor cortex to facilitate response selection based on the learned probability of reward given the current stimulus state. The PMD_{Maint} layer represents possible stimuli to be actively maintained so as to constrain motor selection processes. Its corresponding striatal region learns which stimulus dimensions should be gated into PMd based on the learned probability that their maintenance is predictive of reward. The PMD_{Out} layer represents the deep lamina (e.g., layers 5/6) of PMd in which only a subset of currently maintained PMd stimuli influences response selection, by projecting to the motor striatum. Its corresponding striatal area learns which of the maintained PMd stimuli should be output-gated depending on context. The most anterior prePMd layer maintains stimulus features that act as context, by sending their axons to striatal output-gating areas of PMd. Its corresponding striatal gating layer learns whether the maintenance of particular stimuli as higher order context in prePMd is predictive of reward. Bottom: example network state when presented with color 2, shape 3, and orientation 2. Arrows reflect direct projections, circles reflect BG gating circuitry, and dashed red lines reflect hierarchical flow of control. S3 and O2 are maintained in PMd, and C2 in prePMd. Due to influences of C2 in prePMd, only the shape and not orientation is output-gated. The number of stimulus–response associations is reduced by focusing on PMD_{Out} states.

$$P(\theta_{R,O}|r_1 \ldots r_n) \propto P(r_1 \ldots r_n|\theta_{R,O})P(\theta_{R,O}),$$

where $\theta_{R,O}$ reflects the parameters governing the belief distribution about rewards given the presence of orientation O and the choice of response R, and $r_1 \ldots r_n$ are the rewards observed thus far (in the n trials in which this R was chosen). We then calculated the probability of selecting each of the 3 responses by comparing the expected means μ of their reward distributions via the commonly used softmax logistic function. Thus, the probability of selecting R_i on trial t according to orientation expert O was

$$P^O_{R_i}(t) = \frac{e^{\frac{\mu^O_{R_i}(t)}{\kappa}}}{\sum_j e^{\frac{\mu^O_{R_j}(t)}{\kappa}}},$$

where κ is a noise/exploration–exploitation parameter governing choice stochasticity and was estimated as a free parameter. The same computations were performed in parallel for each expert e, including a shape expert, a color expert, an expert for each of the 2-way conjunctions (shape-orientation, shape-color, and color-orientation),

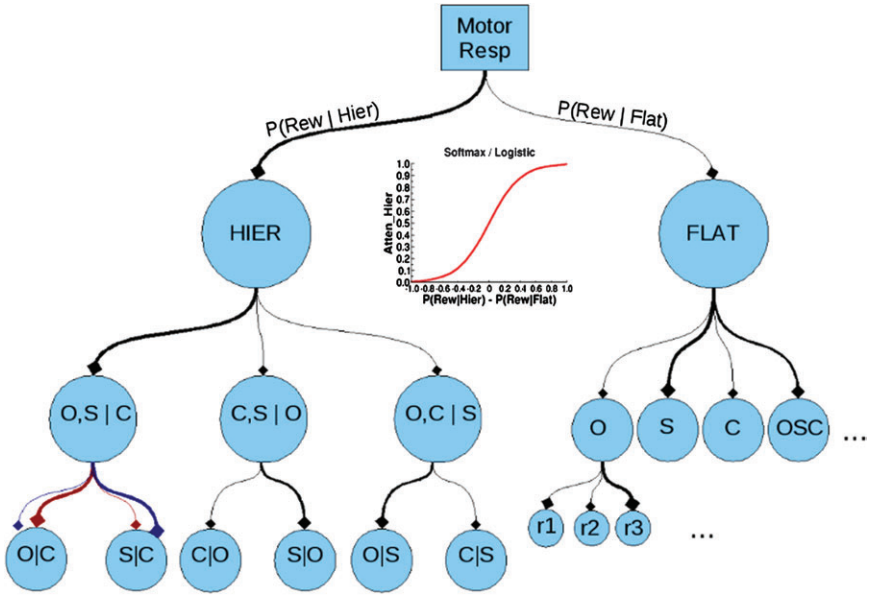

Figure 3. Mixture of experts model. Each flat expert learns reward probabilities for each response given their expert dimension(s) (*O*, orientation; *S*, shape; *C*, color). Responses are selected by each expert using the softmax logistic function. An overall Flat expert learns via a reinforcement credit assignment to allocate attention among the experts in proportion to their reliability. Each hierarchical expert learns to dynamically gate attention to one of 2 dimensions depending on a candidate higher order dimension. The leftmost expert learns to attend to orientation for red contexts and to shape for blue contexts. The overall Hierarchical expert learns which of the hierarchical experts is most reliable, and the overall motor response is selected as a mixture between the two top level experts(representing whether the task structure is likely to be hierarchicalor flat), again in proportion to their reliabilities.

and a full 3-way conjunctive expert which separately learns reward statistics for each response given the specific combination of all 3 stimulus dimensions. In this manner, if any of the stimulus dimensions (or their combination) is reliably indicative of a reinforcing response, then, given sufficient experience, the associated expert will detect this statistical regularity and assign that response a high probability.

Note that the full (3-way) conjunctive expert will eventually learn the optimal response for each combination of stimulus dimensions. However, this learning is inefficient: it depends on having relevant experience for each possible combination of orientation, shape, and color (thus requiring many trials and a high memory capacity) and has no mechanism to generalize learning from any one combination of cues to any other. In contrast, a unidimensional (e.g., orientation) expert, after observing a reward for a given response, will generalize its learning to other instances: when the same orientation appears again in the context of some other shape and/or color, it will immediately assign a greater probability of selecting the same response. Of course, depending on the task structure this may or may not be adaptive. Thus, a key issue is how to allocate attention to each expert in any given trial.

Learning the Reliability of Individual Experts

To maximize positive outcomes, the multiple expert scheme should be able to learn the reliability of each expert's response–reward predictions across trials and to differentially weight (attend to) their assigned response probabilities based on this reliability. In some cases, it is further useful to dynamically gate attention to the experts depending on the current stimulus context, rather than simply according to past overall success (e.g., Jacobs et al. 1991; Kruschke 2001).

We considered various approaches for learning the attentional weights that should be allocated to each expert. From a Bayesian optimality perspective, the attentional weights for each expert can be computed by computing the posterior probability that each expert encompasses the best account of the data:

$$P(e|r)=\frac{P(r|e)P(e)}{\sum_i P(r|e)P(e_i)}.$$

Thus, the evidence for each expert under this scheme is computed by considering its prior evidence and the likelihood that the observed (positive or negative) reward r would have been observed under the expert's model, relative to all other experts. (The likelihood $P(r|e)$ is simply the learned reward probability of the selected response under the expert, as described above.) For example, if there was a low reward probability for the selected response under a given expert, and a negative outcome is observed, then the likelihood of the observation occurring given the expert's model is high. Once the posterior evidence for each expert is computed, one can then apply Bayesian model averaging to allocate attentional weights to each expert in proportion to their log evidence.

Although this scheme for selecting among experts is optimal, in light of the neural model (and others in this domain), we considered an alternative approximation heuristic based on reinforcement learning that may be more related to corticostriatal mechanisms. Here, we represent participants' belief that each expert is predictive of reward with a Beta distribution, in a similar manner as that described for individual responses under each expert, and apply Bayes' rule to learn the probability that its predictions are correct as whole. We make use of a credit assignment mechanism that rewards/punishes experts only if they contributed to the observed outcomes. (See Discussion for posited neural mechanism of this credit assignment). Using the orientation expert as an example, following each outcome we update the posterior distribution as follows:

$$P(\theta_O|r_1'\ldots r_n')\propto P(r_1'\ldots r_n'|\theta_O)P(\theta_O),$$

where r' are the credit assignment–filtered rewards indicating whether the expert contributed to the positive (negative) outcome and should therefore receive credit (blame). Specifically, if R_i is the selected response, rewards are delivered to the expert as follows:

$$r'=\begin{cases} r & \text{if } \mu_{R_i}>\mu_{R_j}, \forall j\neq i \\ 1-r & \text{otherwise} \end{cases}.$$

In words, experts were rewarded only when an actual reward was received ($r=1$) and that expert assigned the largest probability to the executed response (i.e., the expert contributed to the choice). If the expert predicted that an alternative (unselected) response had a higher likelihood of being rewarded that expert was punished ($r'=0$) because it did not contribute to the reward. Conversely, if a reward was not received ($r=0$), then the expert is punished ($r'=0$) if it contributed to

the negative outcome and was optimistically rewarded ($r' = 1$) otherwise (i.e., positive credit is given to an expert that did not contribute to the negative outcome because the response it had assigned the highest reward probability was not selected in that trial). This credit assignment scheme operated at all levels—to the individual flat and hierarchical experts and to the 2 higher level Hierarchical and Flat experts. The rationale is related to the neural model, in which reward prediction errors differentially reinforce striatal Go/NoGo neurons coding for the action/representation that had been gated (e.g., color as higher level context in prePMD) compared with those for other representations that were not gated.

Finally, we also considered (and simulated) more graded credit assignment mechanisms, in which experts are rewarded in proportion to their relative assigned probability that a response is correct (rather than all or none), to estimate a "responsibility signal" in the multiple model-based reinforcement learning (MMRBL) algorithm (Doya et al. 2002). (This framework also allows the learning in each expert to occur in proportion to the attentional weight currently assigned to it). We considered 2 forms of this algorithm (with and without memory decay; see Supplementary Material for full equations). The general results of these simulations were similar to that described above, but they produced weaker fits to the behavioral data than this discrete credit assignment mechanism (which also matches more closely to the mechanism posited in the neural models for credit assignment; see Discussion). In general, our strategy was to find the model that provided the best overall fit to behavior, which would then permit an analysis of the neural correlates of its computations with fMRI data, which is most interpretable when it corresponds to a well-fit model.

Arbitration

We can now differentially weight the contributions of the different experts in proportion to their learned probability of having contributed to rewarding outcomes. To do so, we make use of the same softmax function as that used to arbitrate between responses for a given expert to generate attentional weights for each expert relative to all others. Returning to the orientation expert as an example, the attentional weight w_O in trial t is computed as a function of its assigned expected reward probability μ^O relative to that of all experts:

$$w_O(t) = \frac{e^{\frac{\mu^O(t)}{\zeta}}}{\sum_E e^{\frac{\mu^E(t)}{\zeta}}},$$

where ζ is a gain parameter discriminating between the different experts, akin to the exploitation parameter at the response selection level. (In the Bayesian model averaging scheme, the means of the expected reward probabilities are replaced by the posterior evidence for each expert. Although strictly speaking Bayesian model averaging would involve a linear combination of posterior evidences for each expert rather than the nonlinear softmax function which increases discriminability between experts if one is deemed more likely to be correct. Nevertheless, we estimate free parameter ζ, which effectively determines the degree to which differentiation between experts is nonlinear.) An overall probability of generating response R_i can then be computed by simply mixing the experts E additively (a product of experts can also be employed but does not produce substantially different results) in proportion to their attentional weights:

$$P^f_{R_i}(t) = \sum_E w_E P^E_{R_i}\left(t\right),$$

where P^f refers to the probability of generating responses for an overall "flat expert" combining all of the subordinate experts described thus far. Given any task structure mapping stimuli to responses and rewards, this MoE will eventually learn to produce the probabilistically optimal responses. (If the structure is arbitrary and no particular dimension is more predictive than any other, the full conjunctive expert would eventually learn the appropriate mappings for each case and would attain the highest attentional weight.) Note, however, that by itself this scheme does not have the capability to detect any possible hierarchical or branching structure to the task. We thus refer to the weighted combination of these experts as a "flat" expert.

Hierarchical Experts

To learn about possible hierarchical structure in the Badre et al. (2010) task, we introduce hierarchical experts that learn statistics about 2 of the stimulus dimensions conditional on the identity of a third candidate higher order feature (Fig. 3). Because the learner does not know a priori which, if any, feature is the higher order feature, we allow for multiple possible hierarchical experts. For example, the ultimately "correct" hierarchical expert $h_{OS|C}$ learns response–reward probabilities for both shape and orientation separately for each (higher level) color C. To do so, it makes use of 2 subordinate experts that learn the probability of reward for selecting each response and orientation O (or shape S) given color C:

$$P\left(\theta_{R,O|C}\middle|r_1 \ldots r_n\right) \propto P\left(r_1 \ldots r_n\middle|\theta_{R,O|C}\right) P\left(\theta_{R,O|C}\right),$$

$$P\left(\theta_{R,S|C}\middle|r_1 \ldots r_n\right) \propto P\left(r_1 \ldots r_n\middle|\theta_{R,S|C}\right) P\left(\theta_{R,S|C}\right).$$

Credit assignment operates as above, but now across subordinate experts within the hierarchical expert, allowing this expert to attend to orientation or shape depending on their reliability of predicting reward for each color. Specifically, the hierarchical expert $h_{O,S|C}$ dynamically assigns attentional weights to shape or orientation contingent on the color:

$$w_{O|C}(t) = \frac{e^{\frac{\mu^{O|C}(t)}{\zeta}}}{e^{\frac{\mu^{O|C}(t)}{\zeta}} + e^{\frac{\mu^{S|C}(t)}{\zeta}}},$$

where $w_{O|C}(t)$ is the relative attentional weight to the orientation expert relative to the shape expert when color C is present. The probability of selecting response R_i for this hierarchical expert $h_{OS|C}$ is then simply mixed according to these weights on each trial:

$$P^{h_{OS|C}}_{R_i}\left(t\right) = w_{O|C} P^{O|C}_{R_i}\left(t\right) + w_{S|C} P^{S|C}_{R_i}\left(t\right).$$

We similarly included 2 other hierarchical experts $h_{S,C|O}$ and $h_{O,C|S}$, for which orientation and shape are the higher order features. The credit assignment mechanism is applied to determine, within an overall hierarchical scheme, the probability that each of the 3 hierarchical experts contributes to reward. An overall hierarchical expert assigns attentional weights to these possible hierarchical structures, much like the overall flat expert:

$$P^b_{R_i}\left(t\right) = w_{OS|C} P^{OS|C}_{R_i}\left(t\right) + w_{OC|S} P^{OC|S}_{R_i}\left(t\right) + w_{CS|O} P^{CS|O}_{R_i}\left(t\right).$$

Finally, a second-level attentional selection process was implemented to arbitrate between the overall hierarchical expert and overall flat expert (each of which constituted a weighted combination of subordinate experts)—again based on the learned probability of reward given that these experts contributed to choice. Thus,

$$w_{\mathrm{H}}(t) = \frac{e^{\frac{\mu^{\mathrm{H}}(t)}{\xi}}}{e^{\frac{\mu^{\mathrm{H}}(t)}{\xi}} + e^{\frac{\mu^{\mathrm{F}}(t)}{\xi}}},$$

where ξ determines the gain with which one discriminates between hierarchical and flat structure. The net response selected is then, finally, as follows:

$$P_{R_i}\left(t\right) = w_{\mathrm{H}} P^{\mathrm{H}}_{R_i}\left(t\right) + W_{\mathrm{F}} P^{\mathrm{F}}_{R_i}\left(t\right).$$

Note that each of the hierarchical experts correspond better to a classical "mixture of experts" architecture (e.g., Jacobs et al. 1991; Kruschke 2001) because they dynamically gate the outputs of a given expert conditioned on a context; whereas the overall scheme for combining experts is closer to Bayesian "model averaging" (where we average across individual flat experts and hierarchical MoE).

Quantitative Fits to Human and Neural Circuit Choices

We set out to investigate whether the MoE high level approximation to corticostriatal RL can provide a reasonable fit to individual human participant choices in the Badre et al. task, and similarly, to individual corticostriatal neural networks. A further goal was to use the model-derived attentional weights as a means to investigate neural

computations associated with hierarchical learning in the functional imaging data (see companion paper).

The above goals require inferring the extent to which individuals attend to hierarchical structure, that is, the attentional weights to hierarchical experts in the MoE. By observing each participant's sequences of choices and reward outcomes, we can infer the attentional weights for each expert and how these evolve across the learning phase in hierarchical and flat conditions. Specifically, the attentional weights are found by maximizing the likelihood of each participant's trial-by-trial sequence of choices across all 720 trials, with free parameters adjusted to estimate individual differences and to maximize this likelihood. We describe the nature of these free parameters next.

Although the full MoE structure could, in principle, be considered by each individual (i.e., if all attentional weight distributions are initialized to be uniform), we assumed that participants vary in the extent to which they considered particular hypotheses from the outset and that any particular individual might only consider a subset of the possible structure (e.g., pruning; Daw et al. 2005). We modeled this variation by allowing the prior attentional weights for the different experts to be free. The Bayesian formulation for attentional weights allows us to estimate low priors to certain experts of individuals who are not well characterized as initially attending to these experts such that these priors could be overcome only by overwhelming evidence in their favor (in which case a nonlinear increase in attention could be observed, as is often the case for the hierarchical expert). In contrast, standard reinforcement learning models would have more difficulty capturing these phenomena: low prior values would be rapidly updated as a function of a series of prediction errors. Thus, although formally the Bayesian interpretation, with a low prior for attentional weights, assumes that individuals accrue evidence about experts to which they are not "attending" for a prolonged period, it may simply capture the tendency to initially not attend to a particular feature and then to randomly gate this feature into PFC at some point and only begin learning about it then without any bias.

A critical prior determined the initial likelihood of the participant considering overall hierarchical versus flat structure. Those with high values would be better fit by a model in which they are testing hierarchical structure early on during learning, whereas a very low prior would suggest that they are unlikely to be testing hierarchical structure. Intermediate values would suggest that with sufficient evidence for hierarchy, they may attend to such structure. Of course, individual differences exist in other attentional factors as well. Thus, we included separate free parameters estimating the prior tendencies to attend to unidimensional experts for color, orientation, and shape. We further included a prior for attending to 2-way conjunctions and to the 3-way conjunctive expert. To minimize the number of free parameters, rather than including a separate prior for each 2-way conjunction, we instead estimated a single β hyperparameter for 2-way conjunctions across all 3 of these experts (with high values reflecting a low prior to attend to conjunctions). We then initialized the α hyperparameter for each 2-way expert to reflect the mean prior of the 2 constituent dimensions (with higher values reflecting a tendency to attend to this specific conjunction). (For example, the orientation–shape conjunctive expert would have its prior α set to the mean of the α parameters for the orientation and shape unidimensional experts.)

We also allowed individuals to vary in the softmax gain with which they discriminate between different responses (κ) and between different subordinate experts within overall flat and hierarchical experts (ζ) and between flat and hierarchical experts (ξ). These parameters model the extent to which individuals deterministically choose the responses or experts associated with the current highest predictive reliability, and in principle capture variations in selection functions between distinct corticostriatal circuits (from motor to PMd to prePMd). (Separate parameters allow us to estimate the extent to which individuals attend to hierarchical structure separately from their response selection process within a flat expert.)

A final free parameter is used to model the decay of the hyperparameters characterizing the attentional weights between the 2 blocks (from flat to hierarchical or vice versa). This parameter would be maximum (1.0) if participants assume that what they had learned about attentional weights to different experts in the first block should transfer to the second block (e.g., if shape was deemed relevant in block 1 it would begin with a high attentional weight in block 2, even though the specific shape features would be new). Conversely, this parameter would be near zero if participants assumed that the structure had completely changed (such that the learning process in the second block would proceed according to the same initial priors from the first block).

Model fits were evaluated using log-likelihood of choices under the model, $L=\log(\Pi_t P_{i^*,t})$, where t is trial number and i^*, t denotes the subject's choice on trial t. For each subject, the best-fit parameters are those associated with the maximum L value and are, by definition, the most predictive of the subject's sequence of responses across all trials. We calculated the pseudo-R^2 values, defined as $(L-r)/r$, where r is the log-likelihood of the data under a model of purely random choices ($P = 0.333$ for all trials) (Camerer and Ho 1999; Daw et al. 2006; Frank et al. 2007), and the Akaike's information criterion (AIC) (Akaike 1974), which penalizes model fits with increasing numbers of free parameters so as to favor the most parsimonious model that best fits the data. We also applied a recently described Bayesian model selection method, which evaluates the exceedance probability that a given model is more likely than the others given the full set of AIC's for each model and participant (Stephan et al. 2009). This method is more robust to outliers than a simple comparison of mean AIC fit across the group.

Results

Corticostriatal Circuit Model

We simulated the Badre et al. (2010) hierarchical RL task in the corticostriatal neural circuit model (see Materials and Methods and Supplementary Material). In the hierarchical condition, networks with hierarchical corticostriatal structure—that is, in which prePMd provides contextual input to the output gating units of the PMd—show robust learning curves in the same range as those of human participants. We ran control simulations to demonstrate the potential benefit of hierarchical structure in the network model. First, we disconnected the prePMd convergent inputs to PMd output-gating units while leaving all other parts of the network intact. This simulation allows us to investigate whether the influence of more anterior regions (prePMd) on output-gating of posterior regions (PMd) is adaptive when there is hierarchical structure in the environment. Second, we tested an alternative hierarchical hypothesis in which prePMd units projected to the BG input gating units of PMd, similar to the recent proposal of Reynolds and O'Reilly (2009) in hierarchical working memory tasks.

Hierarchical Structure Improves Learning

Indeed, networks with hierarchical structure perform better than networks without such structure (Fig. 4a). No such benefit of hierarchical structure was found in the flat condition (data not shown) (Networks performed above chance but overall worse in the flat condition [as there are multiple competing stimulus–response mappings to be learned], but the degree of learning was not at all influenced by hierarchical structure.) This result suggests that the hierarchical PFC structure facilitated the transformation of a complex task into a much simpler one but that no advantage of this scheme is found when such a transformation is not possible. Indeed, the qualitative difference in performance between networks with and without hierarchical structure resembles the behavioral difference in performance between the hierarchical and flat conditions.

Dopaminergic Reinforcement of Striatal Gating Is Critical

To demonstrate the impact of dopamine-modulating reinforcement learning in the network, we also ran a batch of networks in which all layers and their connectivity remained intact but

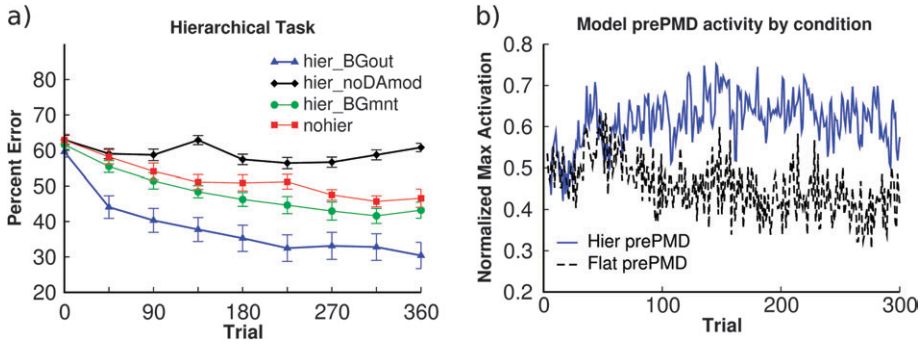

Figure 4. (*a*) Corticostriatal circuit network performance in Badre et al. hierarchical learning task, as a function of trials. Learning is enhanced in hierarchical networks relative to networks with no hierarchical structure (no modulation of PMd circuit by prePMd, "nohier") and relative to networks with hierarchical structure but no dopamine modulation of learning in striatal gating units (hier_noDAmod). (*b*) Activity levels in model prePMd in hierarchical and flat conditions. Results in both panels are averaged across 25 different networks with random initial synaptic weights. Error bars reflect standard error of the mean.

we removed the ability of phasic dopaminergic signals to modulate plasticity in the BG gating units (while preserving learning at the level of motor responses). Thus, the BG can still gate stimuli into the different frontal layers, but gating policies are not adaptively tuned to increase gating of reward-predictive information. In these simulations, frontal representations are not likely to be useful in constraining decisions made at the response level, and indeed, networks exhibited dramatically impaired performance (Fig. 4a). In contrast, intact DA-mediated RL in the gating layers allows the intact network to adaptively gate color dimensional information into the prePMd layer, which then acts as contextual input to the striatal output-gating units for PMd. This output-gating system can then learn which (shape or orientation) dimension is accessed to constrain the motor decision, effectively reducing the number of stimulus–response mappings that the motor system has to learn, thereby making the learning more efficient.

Dynamics of prePMD Activation in Flat and Hierarchical Conditions Matches fMRI Data

Given the finding that hierarchical networks performed best in the hierarchical condition, we then analyzed the activation level in the prePMd layer of the model. Recall that Badre et al. reported that prePMd activation was elevated at the outset of both flat and hierarchical conditions but that the difference between these conditions was reflected by a decrease in prePMd across trials in the flat condition. To examine a possible mechanism for this effect, we plotted the normalized firing rate of the active units in the model prePMd layer as a function of training trials in flat and hierarchical conditions (Fig. 4b). The main difference between the conditions is a decrease in prePMd activity in the flat condition, beginning at approximately trial 50, matching the qualitative pattern found with fMRI (Badre et al. 2010). According to the model, the reason activity declines in this condition is that when no hierarchical structure exists, there is no context (color or otherwise) that reliably predicts when the network should constrain attention to a particular stimulus feature. As such, the prePMd influence can actually hinder performance because it will force the model to focus on a subset of dimensions when it should instead learn about the conjunction of all stimulus features on each trial. Thus, any pattern of prePMd activity elicits a negative reward prediction error, and the resulting "NoGo" learning in the associated BG layer eventually allows it to reduce the probability that stimuli will be gated into (or out of) that layer. As a result, model prePMd activity levels decline with increasing trials in the flat condition. By contrast, in the hierarchical condition, BG gating units are positively reinforced when color is represented in prePMd so that activity is maintained across trials. We confirm this key prediction in the reanalysis of the fMRI data in the companion paper.

Corticostriatal Weights Support Gating Policy Abstraction

To further analyze the mechanisms that support improved performance under hierarchical conditions, we derived an index of gating policy abstraction. Specifically, we computed the summed synaptic connection strengths from the prePMD units representing red (or blue) to the Go and NoGo output gating units in the PMD stripes corresponding to shape (or orientation):

$$hier\_index = \sum \left[w_{C:GO_{S|O}} \right]_+ - \left[w_{C:NOGO_{S|O}} \right]_+ ,$$

where C indicates the relevant presynaptic color unit in prePMD, $S|O$ indicates that weights are computed into post-synaptic striatal units that output gate either shape or orientation (contingent on color identity, according to the hierarchical rule), and $[]_+$ indicates that only weights greater than a threshold of 0.5 are included (lower weights do not tend to contribute to unit activations so are discarded, but this thresholding is not critical). Thus the hierarchical index assesses the degree to which the corticostriatal weights support gating (more Go than NoGo) of the correct hierarchical rule (red:shape and blue:orientation). We also compute the analogous index for the opposite rule that is not supported by the task structure (red:orientation and blue:shape) to ensure that the index is specific to correct hierarchical abstraction and not just increased nondiscriminate propensity for Go gating weights. If the network learns to use hierarchical structure, the weights should evolve to support output gating of the correct dimension and not the other one. Note that this index is a measure of both the tendency for the model to represent color in prePMD (because otherwise these weights from these color units would not evolve) and to correctly map these to the appropriate output gating strategy. These weights also reflect an abstract gating policy because they map onto an entire stripe of shape or orientation, despite the fact that output gating units could also learn to increase weights from the individual instances of each feature (e.g., it could learn specific associations for output gating shape1 vs. shape2).

We found that indeed, with more training the weights evolved to support this abstract output-gating policy, with Go–NoGo gating weights for the hierarchical rule increasing and those for the opposite rule decreasing as learning evolved in the hierarchical block (Fig. 5a). Notably, network accuracy after 360 trials was strongly correlated with this hierarchical index (r = 0.86, P < 0.0001; Fig. 5b). The opposite correlation was seen with Go–NoGo weights to the opposite (incorrect) rule (r = –0.58, P = 0.003), and accuracy was also significantly correlated with relative difference in gating weights between these rules (r = 0.84, P < 0.0001). Similarly, terminal accuracy is correlated with the extent to which the hierarchical index increases from the beginning to the end of the block (r = 0.75, P < 0.0001), demonstrating that this gating policy is learned and not just reflective of random initial weights that happened to support the correct rule in some networks.

Thus, these simulations show that the network learns hierarchical structure. This structure is abstract as it is not tied to any given feature but rather the general tendency for the higher order dimension (color) to increase propensity for gating the lower level dimension (shape/orientation). Thus, once these weights are learned, it is not clear that the "color" units in prePMD should be labeled as such because they now represent which of the other dimensions is relevant (e.g.,

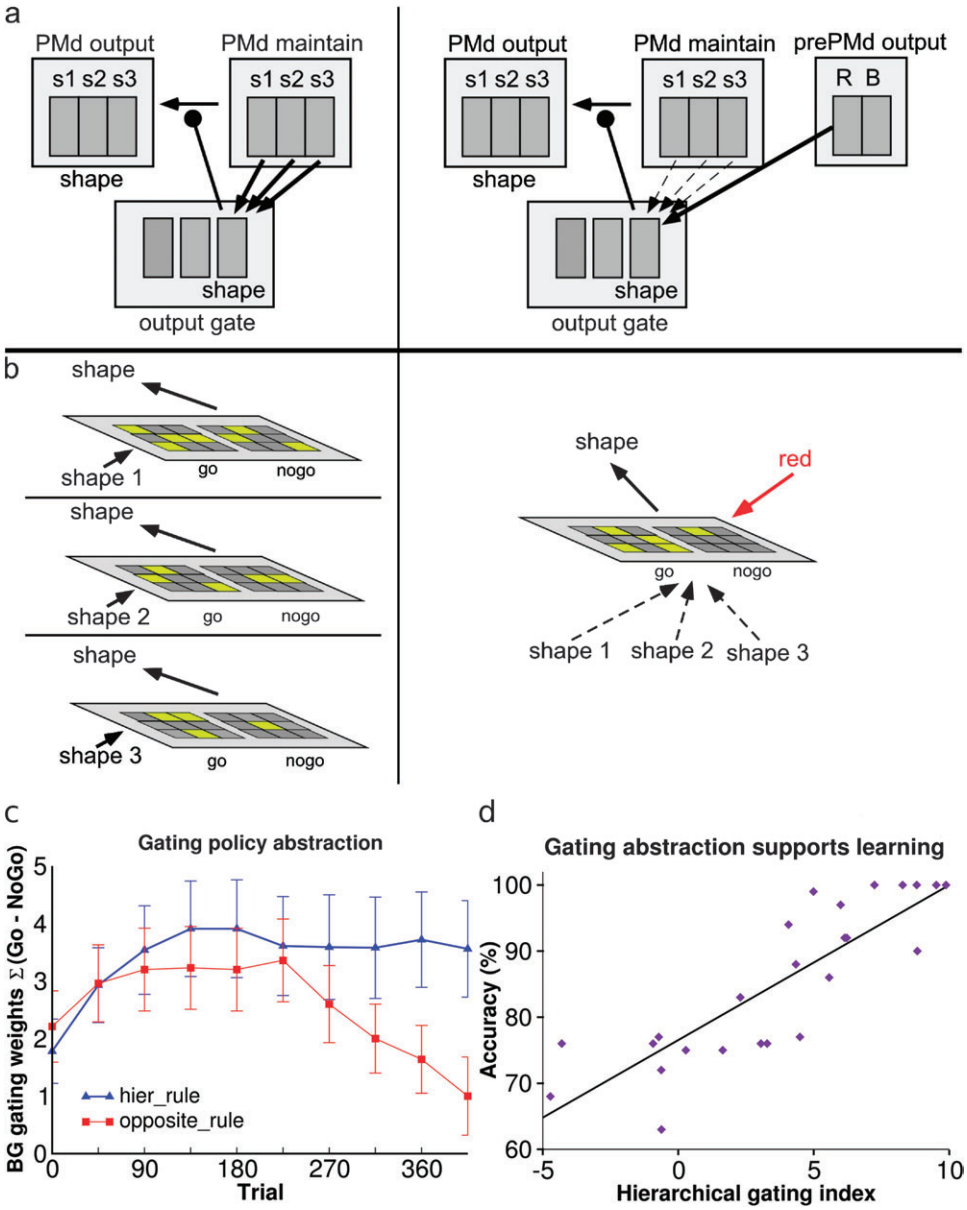

Figure 5. (*a*, *b*) Left: striatal output-gating units from a nonhierarchical network have to learn to output-gate each individual shape feature by assigning strong weights from each the PMd shape units to distinct patterns of Go units. Right: In the hierarchical network, output-gating units of the shape stripe can learn strong Go gating associations whenever the red color unit is active in prePMd. This allows the network to generalize across shapes without learning about each one. (*c*) An index of this hierarchical gating policy abstraction was computed as a function of the weights from prePMd to striatal output gating units in the PMd circuit that support gating of the hierarchical rule (see text). As networks learned in the hierarchical block, the striatum developed an abstract gating policy (e.g., that gates all shapes for a given color, regardless of the particular shape feature), whereas gating weights for the opposite rule decline. (*d*) Across 25 hierarchical networks, the degree of gating policy abstraction at the end of the block was tightly correlated with terminal accuracy.

a prePMD unit, although activated in response to the red input feature, can be said to represent shape). This conclusion extends that of Rougier et al. (2005), who reported that simulated prefrontal units come to represent abstract dimensions (e.g., color) by virtue of maintaining the same activity state over the course of a block in which that dimension was task relevant. Our findings extend this notion such that corticostriatal weights develop to support an abstract hierarchical gating policy that serves to contextualize lower level decisions by higher order features (e.g., units may initially represent color but then come to signal shape). Similarly, a given shape feature unit in PMD initially represents that feature, but with learning comes to signify the appropriate motor response (via its weights to the motor circuit).

In summary, the neural model supports the notion that multiple BG–PFC circuits interact using standard reinforcement learning principles, modulated by dopaminergic prediction errors, to solve hierarchical tasks. Application of this model to a range of other hierarchical (RL and non-RL) tasks is outside the scope of the current study but is currently being investigated. Next, we present results from the abstract model of these processes in individual learners.

MoE Model

As described in the methods, we present results from the best-fitting model to behavior based on Bayesian model selection for group studies. This analysis confirmed that the main MoE model described in the methods is the best-fitting model to the human participant data from the Badre et al. (2010) study (mean pseudo-R^2 = 0.34 and 0.19 for hierarchical and flat blocks, respectively; note these measures apply across all trials including early on when performance and model fit are expected to be at chance levels). Binning the model predictions into bins of width 0.1 from 0 to 1 shows a strong correspondence between predicted probability of a given response and the actual observed allocation of responses for each bin (Fig. 6). Thus, the MoE model provides a reasonable fit to participant choices. Model fits were better when allowing for different softmax parameters for selecting between experts and between motor responses compared with a model assuming a single softmax parameter (exceedance probability P = 0.73). Model fits were also better when including a parameter allowing the attentional priors at beginning of the second block of trials to decay as a function of the posteriors at the end of the first block (exceedance probability P = 0.72). This result suggests participants are more likely than not to test the same structure that they thought described the task in the first block (despite new stimuli). As alluded to in the methods, model fits for our discrete credit assignment mechanism were also greater than that using MMBRL, graded credit assignment, or Bayesian model averaging (exceedance probability for discrete credit assignment P = 0.98). See Supplementary Table for model fit measures.

Attentional Weights

Given these fits to choice data, we next investigated the attentional weights. The attentional weights for a representative participant are plotted in Figure 7 in hierarchical and flat conditions. The bottom of this figure shows attentional weights to the overall hierarchical expert for all participants, showing that individuals differ greatly in the extent to which they appear to attend to hierarchical structure. Some individuals have a relatively

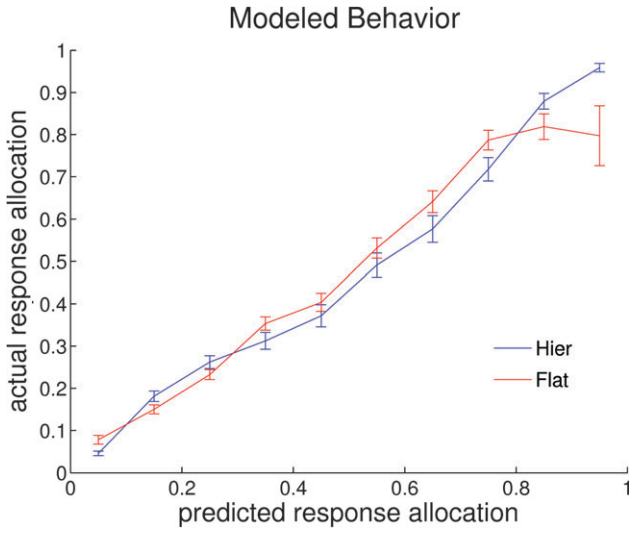

Figure 6. MoE model fits to behavior in Hierarchical and Flat conditions. Graph indicates the relationship between the model's predicted probability that any given response is selected in a given trial (in bins of width 0.1), and the actual proportion of trials in which the associated response was selected by participants in each bin. Results shown across all participants, where each participant's model was optimized by maximizing the likelihood of their trial-by-trial sequence of responses. There was a strong correlation between model predictions and actual choices in both Hierarchical and Flat conditions (r = 0.99 in both cases). Numerically reduced proportion of actual choices in highest bin in Flat condition was associated with a small number of samples for which model predictions were >0.9.

high prior to attend to a particular unidimensional expert (e.g., shape for subject 118), but eventually the reinforcement statistics support a more complex structure (whether hierarchical or not). The attentional weights for the correct hierarchical expert $w_{OS|C}$, relative to all other potential hierarchical structures, increase with experience in the hierarchical block, given the reinforcement statistics. However, individuals differ in the extent to which they attend to the overall hierarchical relative to flat structure (w_H). Some individuals have very low attention to hierarchical structure. To the extent that they perform relatively well in the hierarchical block, these individuals tended to be best characterized by highest attentional weight to the full-conjunctive expert (see supplementary fig. (S109)). Note that even if attention to the conjunctive expert, which can eventually solve the task, is maximal, the learning curve may nevertheless proceed gradually as participants still have to learn the specific stimulus–response mappings for each possible conjunction (and poor memory may can be captured by a low softmax gain selecting among motor responses).

Similarly, even in participants who show rapid increases in attention to hierarchical structure, performance may still take longer to reach asymptotic levels as participants still have to learn specific response weights for individual orientation or shapes. Furthermore, some subjects who eventually show strong attention to hierarchical structure may carry over that belief into the flat block (if it appeared second, as in the example subject in the figure), and this may actually hinder performance given that no hierarchical structure exists in that block. Indeed, across the 10 subjects who performed the hierarchical block first, there was a strong negative correlation between the attentional weights to hierarchical structure (w_H, averaged across the last 50 trials of the hierarchical block) and learning performance in the subsequent flat block (r = –0.74, P = 0.015). This impairment was also observable in just the first

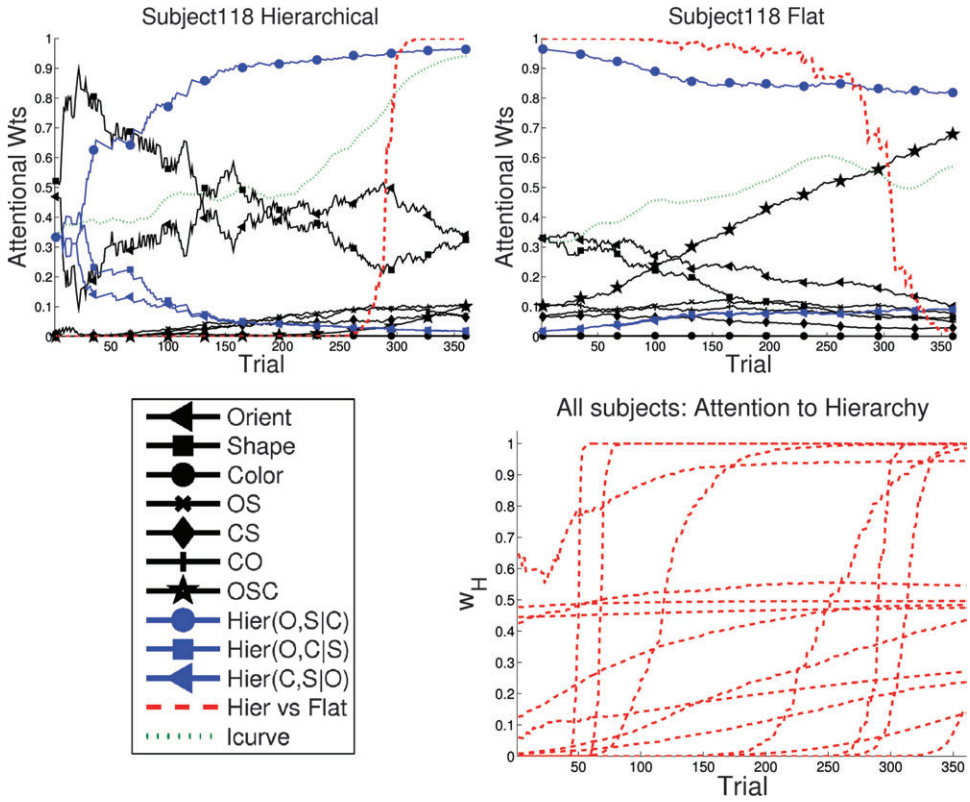

Figure 7. Top: Example attentional weights in a single participant in the hierarchical and flat conditions, as estimated by best fitting model parameters to their trial-by-trial sequences of choices. Within the hierarchical expert, evidence for the correct Hier($O, S|C$) expert increases relatively early on, but the overall attention to Hierarchy relative to Flat (dashed red line, W_H) does not substantially increase until after trial 200. This participant performed the flat condition second and begins with a prior to attend to hierarchy, but when the evidence does not support it, the weight to hierarchy decreases while the eventual winning full conjunctive expert (black asterisk) increases. Green "lcurve" lines reflect smoothed behavioral learning curves as estimated from a Bayesian state space model, which gives probabilistic estimates about the probability of a correct response at each trial (Smith et al. 2004). Bottom: attentional weights to overall hierarchical versus flat expert for all participants. Some participants show rapid increases in attention to hierarchy, whereas others show delayed and/or mixed attention to hierarchy.

50 trials of the flat block ($r = -0.67$, $P = 0.03$), suggesting that it reflects negative transfer. This is a rare example in which participants with "good" learning in one context exhibit poorer learning in another.

Fitting the MoE to the Neural Circuit Model

Overall, the MoE model provides a mechanism by which to infer latent hypotheses tested by individual learners, and its internal variables (e.g., attentional weights) can be used to interrogate neuroimaging data, as we do in the companion paper. Here, we validate the notion that the MoE can infer most likely hypotheses being tested and address its relation to the corticostriatal neural model. To this end, we treated the neural model as a participant, recording its sequence of observations, choices, and rewards and fit the MoE model to these data. The quantitative trial-to-trial fit of choices generated by the neural model was in the same range of that to human data (mean pseudo-R^2 = 0.32, correlations between binned predicted and observed values as in Fig. 6: $r > 0.99$). Thus, in terms of correspondence between observed and predicted data, the MoE model captures the behavioral choices of the neural model to roughly the same degree as it does to human data (despite the fact that clearly the networks, and likely the humans, are not performing exact Bayesian inference).

This exercise allowed us to validate 2 of the assumptions. First, we tested the notion that the MoE model can be used to infer the most likely hypothesis being tested. Because this is a latent (hidden) state, for human participants this is difficult to validate (but see imaging analysis in the companion paper). In contrast, in the neural model, we can directly manipulate whether a particular hypothesis can be tested and then evaluate whether the quantitative fit by the MoE model yields the correct interpretation in terms of its assigned attentional weights. To do so, we ran a batch of networks which were prevented from testing hypotheses associated with an arbitrary stimulus dimension (shape; by disconnecting the shape input units from the frontal areas). We then compared the assigned attentional weights fit by the MoE model to these networks compared with those of the intact networks and found that as expected, attentional weights to the shape expert (W_S) were substantially reduced (near-zero), whereas those to the other experts were unchanged. This analysis confirms that when specific hypotheses are generated to guide action selection, the MoE model can appropriately assign attentional weights to these hypotheses (and that it can do so with the assumptions about the mechanisms of hypothesis testing embedded in the neural model). Second, this same approach allowed us to test the ability of the MoE to infer the likelihood of testing hierarchical structure. We fit behavioral choices of networks with and without hierarchical structure (i.e., by including or excluding projections from prePMd to the striatal output-gating units of PMd). These additional simulations were conducted without working memory demands. That is, stimuli were not removed from the input during the delay period,

allowing networks to potentially solve the task by forming conjunctive associations because all stimulus dimensional information is available on each trial. This allows us to investigate the impact of corticostriatal hierarchical structure on hierarchical attentional weights while ensuring that networks could potentially solve the task with conjunctive mechanisms. Indeed, in these simulations, differences in hierarchical attentional weights emerged at about 150 trials, but a performance advantage in terms of accuracy was not reliably observed until about 300 trials. Both groups of networks showed relatively speeded acquisition under these conditions, performing at approximately 70% accuracy rates by trial 300. Hierarchical networks exhibited another 10% improvement in the last 100 trials, whereas nonhierarchical networks did not. We found that in the hierarchical task condition, the attentional weight assigned to testing hierarchical hypotheses (w_H) was significantly greater in networks with hierarchical structure (Fig. 8). Notably, within hierarchical networks, the estimated w_H was related to the index of gating policy abstraction (see above) in the corticostriatal weights from prePMD to striatal output gating units ($r = 0.47$, $P < 0.02$) during the first half of the block. A similar but nonsignificant trend was observed in the second half of the block ($P = 0.1$). Thus, gating policy abstraction supported the rapid development of hierarchical structure that was revealed in the MoE quantitative fits to network choices. These simulations show that differences in hierarchical strategies can be detected even in the presence of minimal changes in performance accuracy and in advance of such changes.

Note that despite the fact that networks begin the task by gating representations in prePMd, and therefore have the immediate capacity to test hierarchical structure, the attentional weights to such structure as estimated by the MoE often start with low values at the outset of the block. This occurs because 1) the weights from prePMd to striatal output gating units are randomly initialized and it takes experience for these to be tuned such that the output gating policy is reliably influenced by representations in prePMd and 2) networks also have immediate abilities to respond based on unidimensional strategies (e.g., if one or more dimensions is maintained and output-gated from the outset). Thus, the inferred attentional weights to the hierarchical (or any other) expert reflects the tendency to preferentially use this information for guiding responses. Please see the companion paper for more discussion of this issue as it relates to fMRI data in prePMd.

Discussion

We have provided a novel model of hierarchical reinforcement learning in corticostriatal circuits. Existing theories of hierarchical cognitive control suggest a cascade network by which rostral PFC modulates processing in posterior PFC (Koechlin and Summerfield 2007; Badre and D'Esposito 2007; Badre et al. 2009). The mechanisms underlying these asymmetric interactions remain underspecified, and prior work has focused primarily on corticocortical influences in conceptualizing this rostral-to-caudal cascade (Koechlin et al. 2003; Badre and D'Esposito 2007). Our model builds on these prior notions by suggesting that—at least when rules have to be learned—the influence of anterior PFC on posterior PFC may be indirect (rather than directly corticocortical) such that action selection at one corticostriatal level is constrained by inputs from more anterior levels. In other words, hierarchical control may emerge, in part, from multiple nested frontostriatal loops. Reinforcement learning operates at each level such that the system adaptively learns to represent and maintain higher order contextual information in rostral regions (e.g., prePMd), which serve to conditionalize attentional selection in more caudal regions (e.g., PMd), ultimately influencing response selection in motor cortex.

The probabilistic MoE model provides an abstract expression of the key computations of this circuitry and makes quantitative estimates of individual subject's attentional allocation to different features and hypotheses. Together, the models provide key predictions for fMRI data to be tested in the companion paper. Specifically, we test whether estimates of individual differences in attention to hierarchical structure and the modulation by reward prediction errors are associated with differential activations in the corticostriatal circuits associated with second level hierarchical control. We also test whether

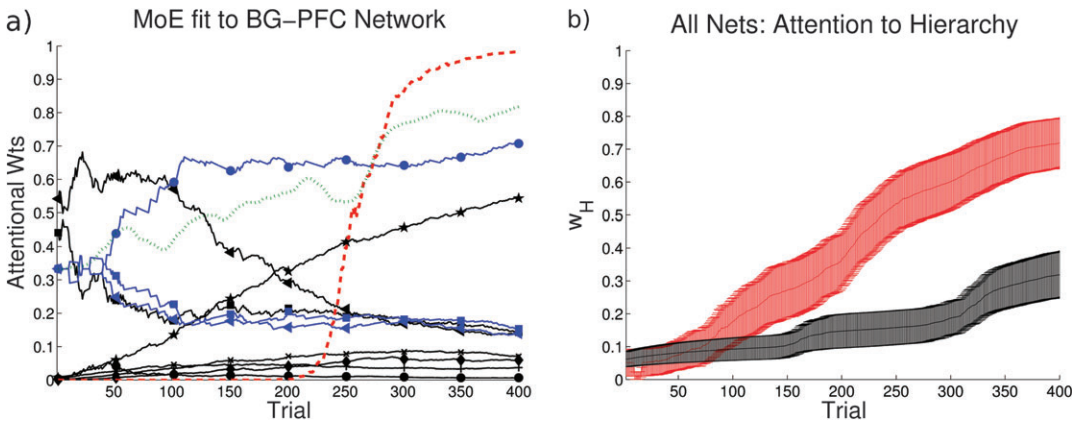

Figure 8. (*a*) Example attentional weights estimated by the MoE fits to trial-by-trial sequence of choices generated by a BG–PFC network in the hierarchical condition. Smoothed learning curve of this network is plotted on the same scale (dotted green line). This network appeared to begin responding primarily relying on unidimensional strategies (particularly orientation; black triangles), which then decrease with experience due to their inconsistent reward associations. The weights to the full 3-way conjunctive expert (black asterisks) increase incrementally as performance improves. Within the hierarchical expert (blue curves), evidence for the correct Hier($O, S|C$) expert (blue circles) increases relatively early on, but the overall attention to Hierarchy relative to Flat (w_H, dashed red line) does not substantially increase until after trial 200. See Figure 7 for full legend indicating the identity of each expert. (*b*) Mean (standard error) attentional weights to hierarchical versus flat expert (w_H) estimated across all 25 networks with (red) and without (black) hierarchical structure (projections from prePMD to output gating units of PMd).

the neural model prediction that the decline in prePMD activity in the flat condition is associated with reward prediction error signaling in the striatum, allowing it to learn not to gate information into (or out of prePMD) when it is not useful.

Relationship between Levels of Modeling

Although modeled at quite a different level of analysis, the development of the computational-level MoE model was motivated by many of the principles embedded within the biologically inspired neural circuit model of corticostriatal interactions. Most obviously, both models learn as a function of positive and negative reward outcomes, which are used to adjust the likelihood of selecting particular actions. In the neural model, the nature of the actions varies from gating of motor responses to the input and output gating of stimulus dimensions for response selection in PMd and finally to the gating of contextual stimuli in prePMd. We represented the different actions in the MoE model by including different levels within the experts, with the lowest level learning about responses and higher levels reflecting attentional weights to different stimulus dimensions. The mechanism for discovering hierarchical structure is similarly related across models. In the neural model, the prePMd provides contextual information to constrain which of the other stimulus dimensions to output gate in order to influence response selection. In the MoE, hierarchical experts dynamically gate attention to one stimulus dimension contingent on the identity of a higher order feature. We showed explicitly that when the MoE is fit to the generative neural model, the estimated attentional weight to the hierarchical expert is proportional to the degree to which the neural model learns an abstract gating policy.

Various neural models of BG show that the corticostriatal circuit adaptively gates one (or a subset) of available alternative actions (Houk and Wise 1995; Gurney, Prescott, and Redgrave 2001; Frank 2006; Humphries et al. 2006; O'Reilly and Frank 2006). In the MoE model, we implemented this selection across all levels of actions with the softmax function, commonly used in algorithmic models of this circuitry (Samejima et al. 2005; Daw et al. 2006; Frank et al. 2007; Doll et al. 2009). Thus, although the models apparently differ in that for any given trial, the neural model stochastically gates stimuli in prePMD and PMD to constrain response selection, whereas the MoE assumes that attentional weights are mixed in each trial. However, this difference is less profound than it appears as the softmax mixture function provides a probabilistic estimate of which response is likely to be selected, rather than a literal mixture of multiple responses. The same applies here in terms of attentional weights to experts. Thus, the main difference here is simply that the neural model is used as a generative model (in which it stochastically gates discrete actions), whereas the MoE is used as a probabilistic estimate of which experts are gated. Simulations validate this claim, showing that the MoE provides a reasonable fit to the neural model's choices.

Finally, in both models, the action-values across all levels are learned via a common reinforcement signal. Moreover, the discrete credit assignment mechanism applied in the MoE to determine which of the experts most reliably contributes to rewards was directly motivated by that applied in the neural models (Frank 2005; O'Reilly and Frank 2006). Specifically, a common issue in neural models is how the system "knows" which of the active neurons to assign credit when a reward occurs. One solution to this problem that substantially improves performance is to allow those striatal Go neurons which actually contributed to the selection of the recently executed action to boost the DA signal that projects back to these same group of neurons. This mechanism is plausible, given that Go cells which inhibit the substantia nigra reticulata would disinhibit a population of DA cells in the substantia nigra compacta, due to inhibitory projections from SNr to SNc (Joel and Weiner 2000; O'Reilly and Frank 2006). In this way, the system can learn to preferentially increase action values of those cells which contributed to the action. Evidence for action-specific prediction error signals in striatum supporting this credit assignment mechanism was recently reported in a human functional imaging study (Gershman et al. 2009).

Despite our attempts to constrain model development and analysis according to principles at the other level of analysis, some differences exist. For example, the models differ in their assumptions about the dynamics of hypothesis testing. In the absence of prior learning, the neural model has a tendency to gate novel stimuli into corresponding PFC regions, allowing it to learn the values associated with these subsequent PFC states and to then reinforce gating signals appropriately (O'Reilly and Frank 2006; Hazy et al. 2010). Here this mechanism also applies to the most anterior prePMd layer such that the model has a bias to gate a stimulus as context and therefore a bias for the prePMd to influence output gating at the level of PMd. In other words, the neural model has a bias to represent hierarchical structure. If such structure is not present in the environment, prePMd activity will not be predictive of reward and the resulting development of anterior striatal NoGo representations prevents further gating. In contrast, while the MoE estimates prior tendencies to attend to hierarchical structure $w_{H,}$ for many subjects this prior is estimated to be low; the MoE allows these individuals to nevertheless eventually discover hierarchical structure if the data support such structure. Despite this apparent difference, quantitative fits with the MoE treating the neural circuit model as a subject revealed that prior attention to hierarchical structure is also estimated to be low in the neural model, despite the fact that we know that prePMd activity is high at the outset of learning in both blocks (as in the fMRI data; Badre et al. 2010). Thus, while prePMd activity supports the representation of higher order features, it does not directly lead to a hierarchical gating policy. Rather, our simulations showed that the more proximal neural index of such a policy is the learned synaptic weights from prePMd to (posterior) striatum. Clearly, prePMD activity is necessary for these weights to develop (according to Hebbian principles), and as such we interpret the prePMD activity as an enabling condition. Of course, it is also possible that the MoE model's estimates of hierarchical attention are overly tied to the reinforcement contingencies associated with attending to such structure and cannot rule out the possibility that other models in which attention is not governed by RL principles may be able to estimate attentional weights that more closely parallel prePMd activity. However, such a model is unlikely to predict the modulation of prePMd and striatal activity by attentional weights and reward prediction error, as seen in the companion paper.

Relation to Other Models

Collectively, these findings support the general notion that multiple controllers can exist in parallel, each subject to

reinforcement learning (e.g., Doya et al. 2002; Holroyd and Coles 2002). In the MMBRL framework, each controller learns as a function of reinforcement, subject to "responsibility signals" that determine which ones were more likely to have contributed to the outcome (Doya et al. 2002). Our models share some of these core features (and indeed, we implemented the MMBRL credit assignment mechanism in an alternative model) but also differ in both conception and implementation. For example, the credit assignment mechanism we adopt relies on discrete output signals indicating whether a given expert contributed to the response in a given trial. This is intended to reflect the hypothesized mechanism in which dopaminergic prediction error signals are selectively amplified in striatal cells that generated the preceding action (Frank 2005; O'Reilly and Frank 2006) and is consistent with recent functional imaging data (Gershman et al. 2009). While this discrete credit assignment mechanism provided a significantly better fit to the behavioral data in this study, further experiments are needed to properly contrast graded versus discrete credit assignment theories.

Our neural model is most closely related to that of Reynolds and O'Reilly (2009); indeed it inherits the same corticostriatal learning and working memory mechanisms as that of its predecessors which did not focus on hierarchical control per se (Frank 2005; O'Reilly and Frank 2006). There are 2 main differences. First, the hierarchical architecture in our model was beneficial for learning of hierarchical structure, whereas that in the Reynolds and O'Reilly model did not improve learning. Second, whereas their model focuses on a role for anterior PFC in contextualizing posterior input gating signals, our model shows that the anterior region (prePMd in this case) is most effective for contextualizing output gating of attentional signals to posterior PFC (PMd), much in the same way that PFC working memory representations are thought to constrain the output gating of motor responses (Frank et al. 2001). However, in other contexts, hierarchical control over both input and output gating functions may be adaptive.

The task simulated here has only minimal working memory demands and is inherently a problem of attentional feature selection in the service of reducing dimensionality of the stimulus–response mappings. Future work will investigate the roles of these features in a broader range of tasks. For example, tasks involving goal–subgoal structure, such as multiple stage arithmetic problems, may benefit from the capacity to maintain overall goals, to constrain attention to the relevant stage of processing, and to update working memory representations as a function of intermediate processing. Consider a task in which participants have to maintain multiple digits in working memory and then perform subsequent computations based on a new instruction (e.g., compute the sum of every odd digit). This example illustrates the need for output-gating: whereas all digits have to be maintained (one does not know until later which will be relevant), only the odd ones should influence decisions. Furthermore, assuming one cannot compute the sum of all odd digits in one step, they might instead consider each digit in memory at a time, decide if it is odd, and if so, add it to the running sum. Thus, the instruction acts as a contextual marker constraining which PFC representations should be output-gated for subsequent operations. (In our example, the contextual information would include not only the constraint that only odd digits should be considered but also whether or not a particular digit had already been added to the sum; this latter constraint may involve yet another layer of hierarchical control.)

How does the model relate to other notions of hierarchical RL? In a recent review, Botvinick et al. (2009) discussed the options framework developed in machine learning as an extension to classical actor-critic RL algorithms. This framework makes a distinction between standard "primitive" actions and hierarchical "options" comprising a string of primitive actions. The top-level state-action policy consists of selecting between primitive actions and options, which have their own policies. Given pretraining of option policies to reach a subgoal, this algorithm is more efficient than standard RL at learning more complex tasks which require obtaining these subgoals in the service of an end goal. On the surface, this domain appears to be quite different than that to which we focus here, where hierarchical policies across the state space are employed on each trial to constrain the selection of a single response. However, we suggest that similar neural mechanisms as explored here may support hierarchical polices across sequences of actions. Consider the case of playing a song on a piano, where the 2 verses involve similar but not identical sequences of keys. We propose that the same architecture could support maintenance of which verse of the song should be played, where in this case the prePMd representation provides the context to the more posterior striatum to output-gate the appropriate sequence of keys. Moreover, once the first verse has been played (i.e., the subgoal has been achieved), the prePMd representation itself has to be updated to set the stage for the next verse and the gating of prePMd may be itself contextualized based on higher order control structures (e.g., DLPFC). Future work will examine this possibility that the maintenance of a longer term goal in more anterior PFC regions across a block of trials can be used to contextualize the output gating of currently relevant subgoals in more posterior regions, which in turn can constrain the selection of primitive actions. Finally, we note that just because sequential action plans can be described hierarchically, this type of hierarchical control need not always be recruited to support them. Indeed, with repeated practice sequential behaviors can be "chunked" into proceduralized action repertoires in the BG (Graybiel 1995). Nevertheless, hierarchical control may be required for branching points in which decisions about alternative action plans need to be contextualized by current plans.

Beyond the models' specific contributions to the domain of learning second-order hierarchical policies, the approach also suggests that computational models can be used to infer latent states in individual learners, by estimating the most likely hypothesis being tested given trial-by-trial observable variables including stimuli, responses, and rewards. The neuroimaging results in the companion paper provide neural evidence for these inferred states by directly interrogating the fMRI data using the quantitative predictions available from the model. Furthermore, we validated the assumption that the MoE model can estimate latent hypotheses by fitting it to choices generated by neural models in which we manipulated their tendency to test specific hypotheses. However, future endeavors using pattern classifiers can be used to test information about the content of distributed neural representations during learning. For example, classifiers can be used to index which dimensions individuals are likely attending to as predicted by the model on a trial-to-trial basis. These and other methods will provide further insight into the neural and cognitive computations associated with hypothesis testing in environments with hierarchical structure and reinforcement learning more generally.

Although our model suggests that the selection between experts occurs via striatal selection mechanisms (including input and output gating) across multiple levels of the hierarchy, it is also possible that the arbitration relies on other cortical mechanisms not included in the model. For example, the medial PFC may act as a monitor to evaluate the success of different experts and to select between them (Samejima and Doya 2007), similar to that proposed by other accounts of medial PFC (Holroyd and Coles 2002). Relatedly, a recent study showed that explicit motivational incentives to perform well at different levels of hierarchical control (contextual vs. episodic) may be mediated by influences of medial PFC (Kouneiher et al. 2009). However, when the task rules have to be learned by trial and error reinforcement, the striatal dopaminergic signals may become relevant. As in other aspects of reinforcement learning, the BG may be important for integrating action-value probabilities over the long run, whereas the medial PFC may monitor rapid changes in reinforcement contingencies and facilitate switching between different strategies on a trial-to-trial basis. Such a dissociation between mediofrontal cortex and striatum in behavioral adaptation in response to changing outcomes on different time scales is consistent with existing theorizing and empirical data (Rushworth et al. 2002; Rushworth et al. 2003; Crone et al. 2006; Frank and Claus 2006; Hampton et al. 2006; Frank et al. 2007; Cavanagh et al. 2010).

In summary, we presented neural and algorithmic models of hierarchical reinforcement learning that provide quantitative fits to human learning and hypothesis testing and the neural mechanisms thereof. We test some of these mechanisms in the companion paper.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

National Institute of Mental Health grant (R01 MH080066-01 to M.J.F.) and NationalInstitute of Neurological Disorders and Stroke grant (NS065046 to D.B.).

Notes

We thank Andy Kaiser, Anne Collins, and Thomas Wiecki for helpful comments and discussion. *Conflict of Interest*: None declared.

References

Akaike H. 1974. A new look at the statistical mode identification. IEEE Trans Automat Contr. 19:716–723.

Alexander G, DeLong M, Strick P. 1986. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci. 9:357–381.

Badre D. 2008. Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cog Sci. 12:193–200.

Badre D, D'Esposito M. 2007. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. J Cogn Neurosci. 19:2082–2099.

Badre D, Hoffman J, Cooney J, D'Esposito M. 2009. Hierarchical cognitive control deficits following damage to the human frontal lobe. Nat Neurosci. 12:515–522.

Badre D, Kayser A, D'Esposito M. 2010. Frontal cortex and the discovery of abstract action rules. Neuron. 66:315–326.

Baier B, Karnath HO, Dieterich M, Birklein F, Heinze C, Muller NG. 2010. Keeping memory clear and stable—the contribution of human basal ganglia and prefrontal cortex to working memory. J Neurosci. 30:9788–9792.

Balleine BW, O'Doherty JP. 2010. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 35:48–69.

Botvinick M, Niv Y, Barto AC. 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition. 113:262–280.

Botvinick MM. 2007. Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. Philos Trans R Soc Lond B Biol Sci. 362:1615–1626.

Botvinick MM. 2008. Hierarchical models of behavior and prefrontal function. Trends Cog Sci. 12:201–208.

Brown J, Bullock D, Grossberg S. 2004. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. Neural Netw. 17:471–510.

Calzavara R, Mailly P, Haber S. 2007. Relationship between the corticostriatal terminals from areas 9 and 46 and those from area 8a dorsal and rostral premotor cortex and area 24c: an anatomical substrate for cognition to action. Eur J Neurosci. 26:2005–2024.

Camerer C, Ho TH. 1999. Experienced-weighted attraction learning in normal form games. Econometrica. 67:827–874.

Cavanagh JF, Frank MJ, Klein TJ, Allen JB. 2010. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. Neuroimage. 49:3198–3209.

Christoff K, Keramatian K, Gordon AM, Smith R, Madler B. 2009. Prefrontal organization of cognitive control according to levels of abstraction. Brain Res. 1286:94–105.

Christoff K, Ream JM, Geddes LPT, Gabrieli JDE. 2003. Evaluating self-generated information: anterior prefrontal contributions to human cognition. Behav Neurosci. 117:1161–1168.

Cools R, Altamirano L, D'Esposito M. 2006. Reversal learning in Parkinson's disease depends on medication status and outcome valence. Neuropsychologia. 44:1663–1673.

Cools R, Barker RA, Sahakian BJ, Robbins TW. 2001. Mechanisms of cognitive set flexibility in Parkinson's disease. Brain. 124:2503–2512.

Cools R, Frank MJ, Gibbs SE, Miyakawa A, Jagust W, D'Esposito M. 2009. Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. J Neurosci. 29:1538.

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

Cools R, Lewis SJG, Clark L, Barker RA, Robbins TW. 2007. L-dopa disrupts activity in the nucleus accumbens during reversal learning in Parkinson's disease. Neuropsychopharmacology. 32:180–189.

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.

Crone EA, Wendelken C, Donohue SE, Bunge SA. 2006. Neural evidence for dissociable components of task-switching. Cereb cortex. 16:475–486.

Dagher A, Robbins T. 2009. Personality addiction dopamine: insights from Parkinson's disease. Neuron. 61:502–510.

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

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

Doll BB, Jacobs WJ, Sanfey AG, Frank MJ. 2009. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res. 1299:74–94.

Doya K, Samejima K, Katagiri K, Kawato M. 2002. Multiple model-based reinforcement learning. Neural Comput. 14:1347–1370.

Draganski B, Kherif F, Klöppel S, Cook PA, Alexander DC, Parker GJM, Deichmann R, Ashburner J, Frackowiak RSJ. 2008. Evidence for

segregated and integrative connectivity patterns in the human basal ganglia. J Neurosci. 28:7143–7152.
Frank MJ. 2005. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and non-medicated parkinsonism. J Cogn Neurosci. 17:51–72.
Frank MJ. 2006. Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw. 19:1120–1136.
Frank MJ, Claus ED. 2006. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol Rev. 113:300–326.
Frank MJ, Fossella JA. 2011. Neurogenetics and pharmacology of learning, motivation, and cognition. Neuropsychopharmacology. 36:133–152.
Frank MJ, Loughry B, O'Reilly RC. 2001. Interactions between the frontal cortex and basal ganglia in working memory: a computational model. Cogn Affect Behav Neurosci. 1:137–160.
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.
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, Seeberger LC, O'Reilly RC. 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 306:1940–1943.
Fuster JM. 1997. The prefrontal cortex: anatomy, physiology and neuropsychology of the frontal lobe. 3rd ed. New York: Lippincott-Raven.
Gershman SJ, Pesaran B, Daw ND. 2009. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J Neurosci. 29:13524–13531.
Graybiel AM. 1995. Building action repertoires: memory and learning functions of the basal ganglia. Curr Opin Neurobiol. 5:733.
Gruber AJ, Dayan P, Gutkin BS, Solla SA. 2006. Dopamine modulation in the basal ganglia locks the gate to working memory. J Comput Neurosci. 20:153–166.
Gurney K, Prescott TJ, Redgrave P. 2001a. A computational model of action selection in the basal ganglia. I. A new functional anatomy. Biol Cybern. 84:401–410.
Gurney K, Prescott TJ, Redgrave P. 2001b. A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. Biol Cybern. 84:411–424.
Haber SN. 2004. The primate basal ganglia: parallel and integrative networks. J Chem Neuroanat. 26:317–330.
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.
Hazy TE, Frank MJ, O'Reilly RC. 2007. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. Philos Trans R Soc B Biol Sci. 362:105–118.
Hazy TE, Frank MJ, O'Reilly RC. 2010. Neural mechanisms of acquired phasic dopamine responses in learning. Neurosci Biobehav Rev. 34:701–720.
Hochreiter S, Schmidhuber J. 1997. Long short term memory. Neural Comput. 9:1735–1780.
Holroyd CB, Coles MGH. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev. 109:679–709.
Houk JC. 2005. Agents of the mind. Biol Cybern. 92:427–437.
Houk JC, Wise SP. 1995. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. Cereb Cortex. 5:95–110.
Humphries MD, Stewart RD, Gurney KN. 2006. A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. J Neurosci. 26:12921–12942.
Inase M, Tokuno H, Nambu A, Akazawa T, Takada M. 1999. Corticostriatal and corticosubthalamic input zones from the presupplementary motor area in the macaque monkey: comparison with the input zones from the supplementary motor area. Brain Res. 833:191–201.
Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. 1991. Adaptive mixtures of local experts. Neural Comput. 3:79–87.
Joel D, Weiner I. 2000. The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience. 96:451.
Koechlin E, Corrado G, Pietrini P, Grafman J. 2000. Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. Proc Natl Acad Sci U S A. 97:7651–7656.
Koechlin E, Hyafil A. 2007. Anterior prefrontal function and the limits of human decision-making. Science. 318:594–598.
Koechlin E, Ody C, Kouneiher F. 2003. The architecture of cognitive control in the human prefrontal cortex. Science. 302:1181–1184.
Koechlin E, Summerfield C. 2007. An information theoretical approach to prefrontal executive function. Trends Cogn Sci. 11:229–235.
Kouneiher F, Charron S, Koechlin E. 2009. Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci. 12:659–669.
Kruschke JK. 2001. Toward a unified model of attention in associative learning. J Math Psychol. 45:812–863.
Lehericy S, Ducros M, Krainik A, Francois C, Vande Moortele P, Ugurbi LK, Kim D. 2004. 3-D diffusion tensor axonal tracking shows distinct SMA and pre-SMA projections to the human striatum. Cereb Cortex. 14:1302.
Lehericy S, Ducros M, Vande Moortele PF, Francois C, Thivard L, Poupon C, Swindale N, Ugurbil K, Kim DS. 2004. Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. Ann Neurol. 55:522–529.
McNab F, Klingberg T. 2008. Prefrontal cortex and basal ganglia control access to working memory. Nat Neurosci. 11:103–107.
Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. Annu Rev Neurosci. 24:167–202.
Mink JW. 1996. The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol. 50:381–425.
Montague PR, Dayan P, Sejnowski TJ. 1997. A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci. 16:1936–1947.
Moustafa AA, Sherman SJ, Frank MJ. 2008. A dopaminergic basis for working memory, learning, and attentional shifting in Parkinson's disease. Neuropsychologia. 46:3144–3156.
O'Reilly RC, Frank MJ. 2006. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. 18:283–328.
Palminteri S, Lebreton M, Worbe Y, Grabli D, Hartmann A, Pessiglione M. 2009. Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. Proc Natl Acad Sci U S A. 106:19179–19184.
Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 442:1042–1045.
Postuma RB, Dagher A. 2006. Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. Cereb Cortex. 16:1508–1521.
Pucak ML, Levitt JB, Lund JS, Lewis DA. 1996. Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. J Comp Neurol. 376:614–630.
Reynolds JN, Wickens JR. 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw. 15:507–521.
Reynolds JR, O'Reilly RC. 2009. Developing PFC representations using reinforcement learning. Cognition. 113:281–292.
Rougier NP, Noelle D, Braver TS, Cohen JD, O'Reilly RC. 2005. Prefrontal cortex and the flexibility of cognitive control: rules without symbols. Proc Natl Acad Sci U S A. 102:7338–7343.
Rushworth MFS, Hadland KA, Gaffan D, Passingham RE. 2003. The effect of cingulate cortex lesions on task switching and working memory. J Cogn Neurosci. 15:338–353.
Rushworth MFS, Hadland KA, Paus T, Sipila PK. 2002. Role of the human medial frontal cortex in task switching: a combined fMRI and TMS study. J Neurophysiol. 87:2577–2592.
Sakai K, Inui T. 2002. A feature-segmentation model of short-term visual memory. Perception. 31:579–589.

Samejima K, Doya K. 2007. Multiple representations of belief states and action values in corticobasal ganglia loops. Ann N Y Acad Sci. 1104:213–228.
Samejima K, Ueda Y, Doya K, Kimura M. 2005. Representation of action-specific reward values in the striatum. Science. 310:1337–1340.
Schonberg T, O'Doherty J, Joel D, Inzelberg R, Segev Y, Daw N. 2010. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. Neuroimage. 49:772–781.
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science. 275:1593.
Shen W, Flajolet M, Greengard P, Surmeier DJ. 2008. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 321(5890):848–851.
Siessmeier T, Kienast T, Wrase J, Larsen JL, Braus DF, Smolka MN, Buchholz HG, Schreckenberger M, Rosch F, Cumming P, et al. 2006. Net influx of plasma 6-18f-fluoro-l-dopa (fdopa) to the ventral striatum correlates with prefrontal processing of affective stimuli. Eur J Neurosci. 24:305–313.
Smith AC, Frank LM, Wirth S, Yanike M, Hu D, Kubota Y, Graybiel AM, Suzuki WA, Brown EN. 2004. Dynamic analysis of learning in behavioral experiments. J Neurosci. 24:447–461.
Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. 2009. Bayesian model selection for group studies. Neuroimage. 46:1004–1017.
Stollstorff M, Foss-Feig J, Cook EH, Stein MA, Gaillard WD, Vaidya CJ. 2010. Neural response to working memory load varies by dopamine transporter genotype in children. Neuroimage. 53:970–977.
Surmeier DJ, Shen W, Day M, Gertler T, Chan S, Tian X, Plotkin JL. 2010. The role of dopamine in modulating the structure and function of striatal circuits. Prog Brain Res. 183:148–167.
Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, Hallett M. 2010. Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. Neuron. 65:135–142.
Voytek B, Knight RT. 2010. Prefrontal cortex and basal ganglia contributions to visual working memory. Proc Natl Acad Sci U S A. 107:18167–18172.
Xu Y. 2002. Limitations of object-based feature encoding in visual short-term memory. J Exp Psychol. 28:458–468.