Electronic Supplementary Material for "Toward an Executive without a Homunculus"

Thomas E. Hazy[1], Michael J. Frank[2], and Randall C. O'Reilly[1]

[1] *Department of Psychology, University of Colorado Boulder, 345 UCB, Boulder, CO 80309, USA*
[2] *Dept of Psychology, Program in Neuroscience, University of Arizona, Tucson AZ, 85721, USA*

This document contains electronic supplementary material for the article *"Toward an Executive without a Homunculus: Computational Models of the Prefrontal Cortex/Basal Ganglia System."*

Article abstract: The prefrontal cortex (PFC) has long been thought to serve as an "executive" that controls the selection of actions, and cognitive functions more generally. However, the mechanistic basis of this executive function has not been clearly specified, often amounting to a homunculus. This paper reviews recent attempts to deconstruct this homunculus by elucidating the precise computational and neural mechanisms underlying the executive functions of the PFC. The overall approach builds upon existing mechanistic models of the basal ganglia and frontal systems known to play a critical role in motor control and action selection, where the basal ganglia provide a "Go" vs. "NoGo" modulation of frontal action representations. In our model, the basal ganglia modulate working memory representations in prefrontal areas, to support more abstract executive functions. We have developed a computational model of this system that is capable of developing human-like performance on working memory and executive control tasks through trial-and-error learning. This learning is based on reinforcement learning mechanisms associated with the midbrain dopaminergic system and its activation via the BG and amygdala. Finally, we briefly describe various empirical tests of this framework.

This document contains three sections intended to be included as an online appendix:

1. Instructions for Downloading PDP++
2. Implementational Details: Description of the Leabra Algorithm
3. Description of the Working Memory Tasks to be Modelled in the Multi-Task (MT) Model

1. Instructions for Downloading and Installing PDP++

All models constructed by the authors and referred to in the paper were created using PDP++, a frequently updated, object-oriented, GUI-enabled version of the Parallel Distributed Architecture originally developed by Rumelhart and McClelland (Rumelhart & McClelland, 1986; McClelland & Rumelhart, 1986). Instructions for downloading, installing and using PDP++ are available at:

http://psych.colorado.edu/ oreilly/PDP++/PDP++.html

†Author for correspondence oreilly@psych.colorado.edu.

2. Implementational Details: Description of the Leabra Algorithm

The model was implemented using the Leabra framework, which is described in detail in O'Reilly and Munakata (2000) and O'Reilly (2001), and summarized here. These same parameters and equations have been used to simulate over 40 different models in O'Reilly and Munakata (2000), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model.

(*a*) *Point Neuron Activation Function*

Leabra uses a *point neuron* activation function that models the electrophysiological properties of real neurons, while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically-based implementation makes it considerably easier to model inhibitory competition, as described below. Further, using this function enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge-building between biology and cognition.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\Delta V_m(t) = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t)) \quad (1)$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a constant $\overline{g_c}$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^\infty = \frac{g_e \overline{g_e}}{g_e \overline{g_e} + g_l \overline{g_l} + g_i \overline{g_i}} \tag{2}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n} \sum_i x_i w_{ij} \tag{3}$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma [V_m(t) - \Theta]_+}\right)} \tag{4}$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $X > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning learning mechanisms (e.g., gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0$, $\sigma = .005$), which reflects the intrinsic processing noise of biological neurons:

$$y_j^*(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} y_j(z - x) dz \tag{5}$$

where x represents the $[V_m(t) - \Theta]_+$ value, and $y_j^*(x)$ is the noise-convolved activation for that value. In the simulation, this function is implemented using a numerical lookup table.

(*b*) *k-Winners-Take-All Inhibition*

Leabra uses a kWTA (k-Winners-Take-All) function to achieve inhibitory competition among units within a layer (area). The kWTA function computes a uniform level of inhibitory current for all units in the layer, such that the $k+1$th most excited unit within a layer is generally below its firing threshold, while the kth is typically above threshold. Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition (O'Reilly & Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

kWTA is computed via a uniform level of inhibitory current for all units in the layer as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta) \tag{6}$$

where $0 < q < 1$ (.25 default used here) is a parameter for setting the inhibition between the upper bound of g_k^Θ and the lower bound of g_{k+1}^Θ. These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^\Theta = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i} \tag{7}$$

where g_e^* is the excitatory net input without the bias weight contribution — this allows the bias weights to override the kWTA constraint.

In the basic version of the kWTA function, which is relatively rigid about the kWTA constraint and is therefore used for output layers, g_k^Θ and g_{k+1}^Θ are set to the threshold inhibition value for the kth and $k+1$th most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold, and the remainder below threshold. For this version, the q parameter is almost always .25, allowing the kth unit to be sufficiently above the inhibitory threshold.

In the *average-based* kWTA version, g_k^Θ is the average g_i^Θ value for the top k most excited units, and g_{k+1}^Θ is the average of g_i^Θ for the remaining $n - k$ units. This version allows for more flexibility in the actual number of units active depending on the nature of the activation distribution in the layer and the value of the q parameter (which is typically .6), and is therefore used for hidden layers.

(*c*) *PVLV Equations*

The PVLV value layers use standard Leabra activation and kWTA dynamics as described above, with the following modifications. They have a three-unit distributed representation of the scalar values they encode, where the units have preferred values of (0, .5, 1). The overall value represented by the layer is the weighted average of the unit's activation times its preferred value, and this decoded average is displayed visually in the first unit in the layer. The activation function of these units is a "noisy" linear function (i.e., without the $x/(x+1)$ nonlinearity, to produce a linear value representation, but still convolved with gaussian noise to soften the threshold, as for the standard units, equation 5), with gain $\gamma = 220$, noise variance $\sigma = .01$, and a lower threshold $\Theta = .17$. The k for kWTA (average based) is 1, and the q value is .9 (instead of the default of .6). These values were obtained by optimizing the match for value represented with varying frequencies of 0-1 reinforcement (e.g., the value should be close to .4 when the layer is trained with 40% 1 values and 60% 0 values). Note that having different units for different values, instead of the typical use of a single unit with linear activations, allows much more complex mappings to be learned. For example, units representing high values can have completely different patterns of weights than those encoding low values, whereas a single unit is constrained by virtue of having one set of weights to have a monotonic mapping onto scalar values.

(i) *Learning Rules*

The PVe layer does not learn, and is always just clamped to reflect any received reward value (r). By default we use a

value of 0 to reflect negative feedback, .5 for no feedback, and 1 for positive feedback (the scale is arbitrary). The PVi layer units (y_j) are trained at every point in time to produce an expectation for the amount of reward that will be received at that time. In the minus phase of a given trial, the units settle to a distributed value representation based on sensory inputs. This results in unit activations y_j^-, and an overall weighted average value across these units denoted PV_i. In the plus phase, the unit activations (y_j^+) are clamped to represent the actual reward r (a.k.a., PV_e). The weights (w_{ij}) into each PVi unit from sending units with plus-phase activations x_i^+, are updated using the delta rule between the two phases of PVi unit activation states:

$$\Delta w_{ij} = \epsilon(y_j^+ - y_j^-)x_i^+ \tag{8}$$

This is equivalent to saying that the US/reward drives a pattern of activation over the PVi units, which then learn to activate this pattern based on sensory inputs.

The LVe and LVi layers learn in much the same way as the PVi layer (equation 8), except that the PV system filters the training of the LV values, such that they only learn from actual reward outcomes (or when reward is expected by the PV system, but is not delivered), and not when no rewards are present or expected. This condition is:

$$\begin{aligned} PV_{filter} \quad = \quad & PV_i < \theta_{min} \vee PV_e < \theta_{min} \vee \\ & PV_i > \theta_{max} \vee PV_e > \theta_{max} \end{aligned} \tag{9}$$

$$\Delta w_i = \begin{cases} \epsilon(y_j^+ - y_j^-)x_i^+ & \text{if } PV_{filter} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where θ_{min} is a lower threshold (.2 by default), below which negative feedback is indicated, and θ_{max} is an upper threshold (.8), above which positive feedback is indicated (otherwise, no feedback is indicated). Biologically, this filtering requires that the LV systems be driven directly by primary rewards (which is reasonable, and required by the basic learning rule anyway), and that they learn from DA dips driven by high PVi expectations of reward that are not met. The only difference between the LVe and LVi systems is the learning rate ϵ, which is .05 for LVe and .001 for LVi. Thus, the inhibitory LVi system serves as a slowly-integrating inhibitory cancellation mechanism for the rapidly adapting excitatory LVe system.

The four PV,LV distributed value representations drive the dopamine layer (VTA/SNc) activations in terms of the difference between the excitatory and inhibitory terms for each. Thus, there is a PV delta and an LV delta:

$$\delta_{pv} = PV_e - PV_i \tag{11}$$

$$\delta_{lv} = LV_e - LV_i \tag{12}$$

With the differences in learning rate between LVe (fast) and LVi (slow), the LV delta signal reflects recent *deviations* from expectations and not the raw expectations themselves, just as the PV delta reflects deviations from expectations about primary reward values. This is essential for learning to converge and stabilize when the network has mastered the task (as the results presented in the paper show). We also impose a minimum value on the LVi term of .1, so that there is always some expectation — this ensures that low LVe learned values result in negative deltas.

These two delta signals need to be combined to provide an overall DA delta value, as reflected in the firing of the VTA and SNc units. One sensible way of doing so is to have the PV system dominate at the time of primary rewards, while the LV system dominates otherwise, using the same PV-based filtering as holds in the LV learning rule (equation 10):

$$\delta = \begin{cases} \delta_{pv} & \text{if } PV_{filter} \\ \delta_{lv} & \text{otherwise} \end{cases} \tag{13}$$

It turns out that a slight variation of this where the LV always contributes works slightly better, and is what is used in this paper:

$$\delta = \delta_{lv} + \begin{cases} \delta_{pv} & \text{if } PV_{filter} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

(ii) *Synaptic Depression of LV Weights*

The weights into the LV units are subject to synaptic depression, which makes them sensitive to *changes* in stimulus inputs, and not to static, persistent activations (Abbott, Varela, Sen, & Nelson, 1997). Each incoming weight has an *effective* weight value w^* that is subject to depression and recovery changes as follows:

$$\Delta w_i^* = R(w_i - w_i^*) - Dx_i w_i \tag{15}$$

where R is the recovery parameter, and D is the depression parameter, and w_i is the asymptotic weight value. For simplicity, we compute these changes at the end of every trial instead of in an online manner, using $R = 1$ and $D = 1$, which produces discrete 1-trial depression and recovery.

3. Description of the Working Memory Tasks to be Modelled in the Multi-Task (MT) Model

(*a*) *Motivation*

An important long-term aim is to apply the PBWM model to a wide range of previously-modeled behavioural and neural phenomena that represent the core principles behind our model. There are several motivations for so doing:

- We think that PBWM begins to approach a complete set of mechanisms for supporting the essential functions of cognitive control. The only reasonable way to test this claim is to apply the model to a wide range of behavioural paradigms. The selected paradigms are ideal in this respect because they tap a range of fundamental aspects of cognitive control (elaborated below), have a considerable amount of empirical data, and should be tractable given the success of existing models.
- The existing models may have relied upon various idiosyncrasies of the specific, more simplified algorithms involved — using one unified model to address a wide range of data greatly reduces this concern. This is a widespread issue with cognitive models — any small set of data can be relatively easily modeled, and may not provide a sufficient test, especially for more complex models. The only way to address this issue is to expand the range of data the model is applied to. We have had considerable experience with this approach, using one unified modeling framework to simulate a wide range of cognitive neuroscience phenomena in O'Reilly and Munakata (2000), and using a single hippocampal model to address a wide range of learning and memory data (O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003; Frank, Rudy, & O'Reilly, 2003; O'Reilly & Norman, 2002).

In what follows, we outline the core tasks to be modelled, emphasising the critical cognitive control functions involved in each case, and summarizing some of the key findings.

(*b*) *Core Paradigms*

(i) *Top-Down Biasing in the Stroop Task*

The Stroop task provides the canonical example of top-down biasing in cognitive control, where robustly maintained PFC representations send supporting activations to task-appropriate processing areas in posterior cortex. Although all of the other task domains also include this top-down biasing function, the Stroop task provides its simplest and most direct test. The basic Stroop effect is that, when presented with the word "red" printed in green ink, top-down support of the color processing pathway from corresponding actively maintained PFC representations is required to override the prepotent response of saying "red" in favour of naming the ink color (green). The importance of this extra PFC support for task-appropriate processing is evidenced by a variety of findings from frontally-damaged patients and neuroimaging (e.g., Cohen & Servan-Schreiber, 1992; Banich, Milham, Atchley, Cohen, Webb, Wszalek, Kramer, Liang, Wright, Shenker, & Magin, 2000). The standard pattern of reaction time (RT) results is shown in Figure 1, where the dominant pathway (word reading) is relatively impervious to interference from the non-dominant one (color naming), but conflict greatly slows the non-dominant pathway. PFC impairments differentially affect

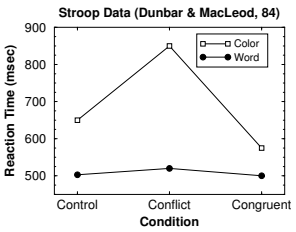

Figure 1. Typical reaction time data from the Stroop task in young healthy adults.

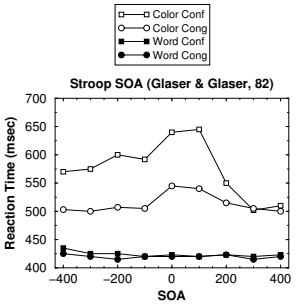

Figure 2. Reaction time data from a Stroop task where colors and words were presented at different times relative to each other (stimulus onset asynchrony, SOA). For word reading negative SOA means the color preceded the word. For color naming, negative SOA means the word preceded the color. Data from Glaser & Glaser, 1982.

this non-dominant conflict condition (e.g., Cohen & Servan-Schreiber, 1992). This basic pattern has been simulated by a number of models based on the top-down biasing principle (Cohen, Dunbar, & McClelland, 1990; Cohen & Huston, 1994; O'Reilly & Munakata, 2000; Herd, Banich, & O'Reilly, submitted).

A more challenging pattern of data comes from manipulating the relative stimulus onset asynchrony (SOA) of the color and word stimuli (Glaser & Glaser, 1982), which has yet to be successfully modeled. The principal challenge here is that presentation of the word prior to the color (negative SOA in the color naming condition) actually produces *less* interference than simultaneous or slightly positive SOA conditions. We plan to address this pattern of results by exploring the effects of stimulus transients on PFC task activations in the PBWM model — prior Stroop models have externally "clamped" the PFC representations and thus lacked any dynamics to the PFC representations.

We have also applied a Stroop model to some initially counter-intuitive fMRI results that seemed to contradict the basic premise of top-down biasing (Herd et al., submitted; Banich et al., 2000). These results showed greater BOLD activation for areas representing the to-be-ignored information (i.e., word reading) in the color-naming conflict condition relative to a neutral color-naming condition, and no difference in activation for task-relevant (i.e., color) processing areas across these conditions. This is just the opposite of what the top-down biasing account would predict (greater activation from the top-down biasing of task-relevant areas, and no effect or even inhibition in to-be-ignored areas). Nevertheless, we showed that the observed pattern of activations is in fact consistent with the top-down biasing framework, as long as semantically-appropriate patterns of connectivity between posterior cortical areas, and more distributed representations in PFC are included in the

model (Herd et al., submitted). Fitting this pattern of results will provide additional important constraints on the model.

(ii) *Working Memory Updating in the AX-CPT Task*

The AX-CPT (A-X version of the continuous performance task, as described earlier in the context of the 1-2-AX task) requires the continuous updating of information in working memory. To correctly discriminate the target A-X sequence from the other possible stimulus sequences (A-Y, B-X, B-Y), the participant must make use of some form of memory for the *cue* stimulus (A or B) in processing the *probe* (X or Y; note that B and Y in general represent a class of non-A,X stimuli). Because of the rapid updating and relatively short duration of these memory demands, it makes sense that dynamically-gated, PFC-mediated working memory would be critical for this task. Indeed, considerable evidence supports this idea (Servan-Schreiber, Cohen, & Steingard, 1997; Braver, Barch, & Cohen, 1999; Cohen, Barch, Carter, & Servan-Schreiber, 1999; Barch, Carter, Braver, MacDonald, Noll, & Cohen, 2001; Braver, Barch, & Cohen, submitted).

The AX-CPT is particularly informative when run with the target A-X sequence presented 70% of the time, and two different interstimulus intervals (ISI) are used (short, 1-2 seconds, and long, 5-8 seconds). The frequency manipulation produces a prepotent bias for pressing the target (right) key when seeing the X, and for expecting an X after seeing an A. The ISI manipulation should affect the strength of the working memory representation of the cue stimulus (A or B). The strength of this cue memory then interacts with the prepotent biases to produce an opposing pattern of performance across the B-X and A-Y trials. A strong memory of the B cue will serve to inhibit the prepotent bias to respond to the X, improving overall performance on B-X, while a weak cue memory will have the opposite effects. Conversely, a strong memory of the A cue will activate the prepotent target-response bias (expecting an X), decreasing performance on the A-Y trials (and a weak memory again has opposite effects). This interaction effect provides a nice control for overall degradation of performance, e.g., in patient populations.

Across a number of studies, patients with schizophrenia exhibited a pattern of responding consistent with impaired working memory for the cue stimulus. Specifically, between the short and long conditions, they exhibited an increase in B-X errors and a decrease in A-Y errors (Servan-Schreiber et al., 1997; Braver et al., 1999; Cohen et al., 1999; Barch et al., 2001). In contrast, healthy control subjects exhibited a slight but statistically significant increase in cue memory related performance across delay: their B-X errors went down while their A-Y errors went up. Furthermore, these effects were specifically localized to the dorsolateral PFC (Barch et al., 2001).

This complex pattern of results across conditions provides a highly constraining dataset for computational modeling (e.g., Braver & Cohen, 2000; Braver et al., submitted). These existing models are based on a dopamine-based dynamic gating mechanism driven by errors in reward expectations according to the temporal-differences (TD) algorithm. In the PBWM model, this TD-driven dopamine mechanism does not directly drive gating — instead, it drives learning of the gating signal produced by the matrix striatal neurons. These TD-driven gating models have also been applied in several other domains discussed below — an important subgoal of the proposed research will be to assess the functional differences between these two types of dynamic gating mechanisms. As discussed next, there are reasons to believe a dopamine-based gating mechanism is insufficient (it does not provide a selective gating signal); we anticipate that other behavioural signatures of the two mechanisms will become apparent in the proposed work.

(iii) *Selective Working Memory Updating and Robust Maintenance in the 1-2-AX Task*

As discussed earlier, the 1-2-AX task extends the AX-CPT task by including an outer-loop of task-demand stimuli (1 or 2) that determine the target sequence operative over a series of inner-loop stimuli. In addition to the rapid updating of the AX-CPT task, this extension requires selective updating (e.g., maintaining the outer loop while updating the inner loop) and places stronger demands on robust maintenance because the outer loop information must be maintained in the face of extensive inner-loop processing and updating. The selective gating demand renders a simple global gating signal (e.g., as provided by a dopamine-based mechanism) insufficient — such a global signal would cause updating of the outer loop every time the inner loop is updated. Other tasks below, when implemented in a more detailed, surface-valid manner, will also likely require selective updating.

To this point, the biggest challenge with this data is simply getting a model to learn to perform such a difficult task — this has been demonstrated with the basic PBWM model (O'Reilly, in press), but we will need to ensure this is still true in the context of learning all the other tasks. The available human data mainly shows that different regions of PFC are used to maintain inner and outer loops (Kroger, Nystrom, O'Reilly, Noelle, Braver, & Cohen, in preparation) — this was predicted from the model. This kind of representational organization is critical to specific aim 2, and is discussed in greater detail there. We also hope to generate testable behavioural predictions by exploring various manipulations in this task further in the model.

(iv) *Basic Task Switching in the Wisconsin Card Sort Task (WCST)*

The WCST is the paradigmatic example of the task-switching flexibility imparted by the PFC. It involves sorting cards having symbols varying in shape, color, and size into corresponding target piles according to one of these three dimensions. The dimensional sorting rule is not told to the participant, who must infer it based on feedback. After a criterion number of successful sorts, the rule is changed (again unbeknownst to the participant) and the number of perseverative sorts according to the old rule is taken as a measure of task-switching flexibility (the fewer the better). People with frontal damage typically perseverate more than healthy controls e.g., Grant & Berg, 1948; Milner, 1963; Heaton, Chelune, Talley, Kay, & Curtiss, 1993, and three-year old children also perseverate more than their four-year old counterparts on a version of the task adapted for children (e.g., Zelazo, Frye, & Rapus, 1996; Munakata & Yerys, 2001).

A number of models have captured performance on these task-switching tasks (Rougier & O'Reilly, 2002; Morton & Munakata, 2002; Dehaene & Changeux, 1991; Levine & Prueitt, 1989). In our model, the critical mechanism was a simple dopamine-based (TD) dynamic gating mechanism that rapidly destabilized previously-maintained task-relevant PFC representations and allowed a trial-and-error search for new such representations to take place when the

rule changed. The speed of this working memory updating in the PFC component was faster than the slow accumulation of weight changes in the posterior cortical component of the model, meaning that the model with a damaged PFC component exhibited more perseveration than the intact network. We expect a similar mechanism to hold for the PBWM model; as with the AX-CPT models, the difference in gating mechanisms will be an important focus.

An aspect of the WCST that remains a challenge for the proposed work is capturing the detailed cognitive operations at work within a given trial as run with human subjects. In our previous models, we have captured the essence of the task by having the model select one out of the possible set of dimensions from a given stimulus at a time. However, people are required to actually place each card on the appropriate matching target card pile — we propose that modeling this process will require a temporally-extended action sequence with an embedded working memory updating step (i.e., requiring selective updating).

(v) *Complex Task Switching in the ID/ED Task*

The intradimensional/extradimensional (ID/ED) task-switching task extends the WCST in a number of interesting ways. It allows for switching to occur at two different levels (within and between stimulus dimensions) and in two different ways (a shift to new stimuli, or reversal of existing stimuli). Relevant data exists across studies with monkeys, neurologically intact humans, frontal patients, Parkinson's patients, and Huntington's patients (Rogers, Andrews, Grasby, Brooks, & Robbins, 2000; Dias, Robbins, & Roberts, 1997; Owen, Roberts, Hodges, Summers, Polkey, & Robbins, 1993; Roberts, Robbins, & Everitt, 1988). The potential to compare the roles of the PFC and BG across frontal and Parkinson's patients should be particularly informative for the PBWM model.

We have already modeled the data from monkey's with lesions to either ventromedial or dorsolateral areas of PFC (O'Reilly, Noelle, Braver, & Cohen, 2002; Dias et al., 1997). This data showed that ventromedial (orbital) lesions produced selective deficits in intradimensional reversals (IDR), while dorsolateral lesions produced selective deficits in extradimensional shifts (EDS). We accounted for this pattern within the top-down biasing framework in terms of ventromedial areas representing detailed, featural information about the stimuli, while dorsolateral areas encoded more abstract dimensional representations. Thus, switching within dimensions was facilitated by the more feature-specific ventromedial representations, and switching between dimensions was facilitated by the more abstract dorsolateral representations. In both cases, the underlying mechanism of rapid updating in PFC, supported by the dynamic gating mechanism, was critical.

In addition to replicating our existing model, we propose to address the range of data observed across the various patient populations studied in the ID/ED task. One consistent finding across these studies is that extradimensional shifting (EDS) is impaired by PFC damage in humans, but intradimensional reversals (IDR) are not reliably impaired. However, both EDS and IDR are impaired in unmedicated Parkinson's patients (Owen et al., 1993), while early-stage Huntington's patients primarily exhibit difficulties with EDS and not IDR. Meanwhile, neuroimaging studies show PFC activation for EDS but not IDR, which nevertheless activates the basal ganglia (Rogers et al., 2000).

Broadly speaking, these results show that the human PFC and BG are both important for task switching, consistent with our framework. The EDS results are directly in line with our existing model and should not pose a difficulty for the PBWM model. The IDR results are more complex, suggesting on balance a role for the BG, but not the PFC. In keeping with our existing model, we wonder if perhaps the regions of the PFC that maintain detailed, feature-level information have not been appropriately targeted in the frontal patient studies (e.g., these regions might be coextensive with relatively posterior and lateral PFC areas involved in language function — such patients typically exhibit dense aphasia and cannot be easily tested in these types of experiments). Consistent with this suggestion, the neuroimaging study did find elevated blood flow in BA9/10, but these results did not remain statistically significant after correcting for multiple comparisons (Rogers et al., 2000). In any case, these studies provide a rich dataset with which to test our model.

(vi) *Dynamics of Perceptual Attention in the Eriksen Flanker Task*

The Eriksen flanker task provides a test of the dynamics of interaction between maintained PFC representations and attentional processing in the posterior cortex. In this task, participants see displays containing a central target item flanked by distractor stimuli, which are either consistent or inconsistent with the target (e.g., SSSSS, SSHSS, HHSHH, HHHHH) (Eriksen & Eriksen, 1974). The simple goal of the task is to identify the target item. Nevertheless, the distractors influence processing, as evidenced by faster RT's for consistent displays and slower RT's for inconsistent ones. Furthermore, when RT's are obtained at a range of durations from short to long, accuracy in conflict conditions initially goes below chance before rising to relatively high levels. This indicates that the flankers initially have more control over performance than the target, but this is overcome through top-down biasing from the PFC for the central target location. Thus, this task enables the time-course of top-down influence to be measured, and fitting this data with the model will provide important constraints on the dynamics of attentional processing between PFC and posterior cortex.

In addition, the Eriksen task has been used to assess the feedback of conflict monitoring systems (in the anterior cingulate cortex) on performance on subsequent trials (e.g., Carter, Braver, Barch, Botvinick, Noll, & Cohen, 1998; Botvinick, Nystrom, Fissel, Carter, & Cohen, 1999; Botvinick, Braver, Barch, Carter, & Cohen, 2001). Our basic PBWM model does not include this mechanism, but we plan to incorporate it pending initial success in simulating single-trial performance across the range of tasks. Thus, this task straddles the core and extension phases.

(c) *Paradigms to be Addressed in the Extended Models*

The following task paradigms will require extensions of the core model, and are therefore to be addressed after an initial model of the core phenomena has been developed.

(i) *Familiarity vs. Active Maintenance in the ABCA/ABBA Task*

A given memory function can often be achieved in a number of different ways in the brain, depending on the specific task demands. In all the core paradigms, it is clear that the

PFC/BG system is required. However, there is good reason to believe that using the PFC/BG system is metabolically costly relative to other forms of memory, and that its use is subjectively experienced as effortful. Therefore, it would be useful to explore whether a model can learn to use this system only when necessary, relying on simpler, less costly memory systems when possible, because this would likely provide an explanation for a variety of behavioural findings.

We believe the two fundamental mechanisms for memory are weight changes and maintained activations, and each has its own strengths and limitations (e.g., weight changes are enduring, but not as flexible, while activations are transient but more flexible; O'Reilly & Munakata, 2000). Generally speaking, we assume that transient, rapidly updated information relies on activation-based maintenance, for which the PFC/BG system is specialised. Nevertheless, there are some such situations in which a weight-based signal can provide a useful basis for task performance.

One context where this tradeoff has been explored is in a variation of the widely-studied delayed-match-to-sample task known as the ABCA/ABBA task (Miller, Erickson, & Desimone, 1996; Miller & Desimone, 1994). Here, monkeys were presented with a sequence of stimuli, and trained to respond whenever the first stimulus repeated. Thus, ABCA refers to a sequence of stimuli, where the final A is the repeat of the first stimulus. Miller and colleagues expected prefrontal working memory to be used to maintain this A stimulus over the duration of intervening items, but their neural recordings actually revealed that this task was being solved via a weight-based familiarity signal encoded in inferior-temporal (IT) cortex. However, when they ran the ABBA version of the task, which contains an embedded repeat of the B stimulus, this familiarity-based system was fooled (the second B was familiar and thus triggered responding, even though it wasn't the first item). After some amount of additional training, the monkeys succeeded in solving the ABBA task, but this time by using actively maintained PFC representations.

In addition to simulating this basic strategy shift, there are detailed patterns of PFC and IT neural firing that can be used to constrain and test the model. We have already developed a model of how a single exposure to an item can produce a robust familiarity signal in an IT-like network (Norman & O'Reilly, 2003). This familiarity effect falls out of the basic Leabra learning and activation mechanisms, but requires a specific monitoring process to reliably detect the global familiarity signal.

(ii) *Higher Capacity, Generalisable Working Memory in the Phonological Loop*

It is likely that a substantial amount of everyday working memory function derives from a specialised system for actively maintaining verbal information called the "phonological loop" (e.g., Baddeley, 1986; Baddeley, Gathercole, & Papagno, 1998; Burgess & Hitch, 1999). This system is thought to be capable of sequentially refreshing a number of verbal (phonological) representations to maintain information in an active state (e.g., somewhat like a continuously repeating tape loop). When one encodes a phone number in working memory, or someone's name, or a relatively short list of words in a laboratory study, the phonological loop is at work. Neuroimaging and patient data show that this system involves the PFC around Broca's area, and a posterior cortical area (supramarginal gyrus) thought to represent phonological content (e.g., Paulesu, Frith, & Frackowiak, 1993; Shallice & Vallar, 1990).

Because verbal information is so overtrained, and there are a finite number of phonemes that can be combined in so many different ways, the phonological loop can have higher capacity and greater flexibility for maintaining novel information than other forms of working memory (O'Reilly & Soto, 2002). Therefore, it represents an important target for our model. However, the required overtraining on a large space of language material is computationally demanding, and requires that we initially develop this model separate from the core model. Therefore, we propose to build on our initial phonological loop model (O'Reilly & Soto, 2002) to address a number of phenomena listed below. After this, we will attempt to integrate this model with our core model, and explore the role of the phonological loop in the core task paradigms, and others described below.

Some of the signature data on the phonological loop include (see Burgess & Hitch, 1999 for full details): a) word list capacity is a function of word length, phonemic similarity, word familiarity (and phonemic similarity to known words for nonwords) and is affected by articulatory suppression; b) serial position effects, including primacy and recency effects; c) interactions between serial position effects and presentation modality and phonemic similarity; d) temporal grouping improves sequential order memory; e) the preponderance of errors are in sequential ordering, not in the items themselves; and f) items from previous lists intrude at the same position in the current list. This is clearly a highly constraining set of data, and the existing model of Burgess and Hitch (1999) provides a clear standard of comparison for our own efforts.

Our existing model of the phonological loop is based on the principles behind the PBWM model, but with some of the difficult learning work achieved by hand-coding because the learning mechanisms had yet to be fully developed (O'Reilly & Soto, 2002). This model demonstrated how the highly-practiced nature of verbal short term memory, combined with the limited number of different phonemes, can enable this system to generalise well to novel phonological sequences (e.g., over 90% correct performance on novel sequences after training on less than 20% of the space). Furthermore, we argued that these same properties should give this system a significantly larger capacity than basic limit of 3-4 items (Cowan, 2001). This model differs from the Burgess and Hitch (1999) model by not requiring any transient Hebbian associations during encoding — this gives it more flexibility, but it remains to be seen how these differences will impact the model's ability to address the data listed above.

(iii) *Working Memory Scanning in the Sternberg Task*

The classic Sternberg memory scanning task explores how people access items maintained in working memory. Each trial of the task involves presenting a variably-sized set of items to be remembered, followed by a target item, to which the participant is to respond Yes if the target was in the memory set (positive case) and No otherwise (negative case). The remarkable finding from this task is that reaction times for both positive and negative trials scale equally and linearly with the memory set size (each additional item requires approximately 40msec of additional time) (Sternberg, 1966). This suggests a process where people sequentially scan memory items, using an exhaustive (non-self-terminating) search for both positive and negative cases.

This is surprising, as one might expect search to terminate when a positive case is found, as is typical with visual search paradigms.

This task has not to our knowledge been simulated in a neurally-based model, and it presents a number of important challenges. Simulating the process of sequential comparison of working memory items with a target item, based strictly on internally-driven operations, will place important demands on the PBWM model. Our initial interpretation of this task is that it depends on the phonological loop. The sequential process matches the sequential search process observed in the Sternberg task quite well, and the highly overtrained nature of this system may explain why it cannot be interrupted after locating a match. Thus, we propose to simulate the Sternberg task using our phonological-loop augmented model.

(iv) *Continuous Updating of Sequential Position Information in the n-Back Task*

The n-back task has become one of the most widely used measures of working memory function, in part because it is a difficult task that readily activates the PFC in neuroimaging studies (e.g., Braver, Cohen, Nystrom, Jonides, Smith, & Noll, 1997; Cohen, Perlstein, Braver, Nystrom, Noll, Jonides, & Smith, 1997; Smith & Jonides, 1998). Like the AX-CPT and 1-2-AX tasks, the n-back involves the sequential presentation of stimulus items. The participant must detect a repetition of any stimulus that occurred n steps earlier in the sequence (where n can be varied from 1-5; most people cannot perform well above 3). The 1-back is relatively simple, requiring only the detection of an immediate repetition. In the 2-back, the preceding 2 stimuli must be maintained, and continuously updated with each new stimulus, while also checking for the repetition. For example, if the sequence starts out with MQABA, the 2nd A is a target. If a B appeared next, it would also be a target, and so on. This simultaneous updating of memory for recent items (which must be maintained in sequential order) combined with the repetition match checking, places strong demands on the working memory system.

Like the Sternberg task (with which the n-back shares a memory scanning component to detect repetition), the n-back task likely loads heavily on the phonological loop system (which is consistent with the neuroimaging data showing Broca's area activations; Braver et al., 1997; Cohen et al., 1997; Smith & Jonides, 1998). The sequential ordering capabilities of the phonological loop are particularly important. Therefore, we propose to simulate the n-back using our phonological-loop augmented model — just getting this model to perform this difficult task will be a monumental achievement.

4. References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*, 220.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*, 158.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Banich, M. T., Milham, M. P., Atchley, R., Cohen, N. J., Webb, A., Wszalek, T., Kramer, A. F., Liang, Z. P., Wright, A., Shenker, J., & Magin, R. (2000). fMRI studies of Stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. *Journal of Cognitive Neuroscience*, *12*, 988–1000.

Barch, D. M., Carter, C. S., Braver, T. S. ans Sabb, F. W., MacDonald, A. r., Noll, D. C., & Cohen, J. D. (2001). Selective deficits in prefrontal cortex function in medication-naive patients with schizophrenia. *Archives of General Psychiatry*, *58*, 280–8.

Botvinick, M., Nystrom, L. E., Fissel, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*, 179–181.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Braver, T. S., Barch, D. M., & Cohen, J. D. (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, *46*, 312–328.

Braver, T. S., Barch, D. M., & Cohen, J. D. (submitted). Mechanisms of cognitive control: Active memory, inhibition, and the prefrontal cortex.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell, & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of frontal cortex involvement in human working memory. *NeuroImage*, *5*, 49–62.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D. C., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*, 747–749.

Cohen, J. D., Barch, D. M., Carter, C. S., & Servan-Schreiber, D. (1999). Schizophrenic deficits in the processing of context: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, *108*, 120–133.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*(3), 332–361.

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In C. Umilta, & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 1–19). Cambridge, MA: MIT Press.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activity during a working memory task. *Nature*, *386*, 604–608.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior

and biology in schizophrenia. *Psychological Review*, *99*, 45–77.

Cowan, N. (2001). The magical number 4 in short-term memory; a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Dehaene, S., & Changeux, J. P. (1991). The Wisconsin Card Sorting Test: Theoretical analysis and modeling in a neuronal network. *Cerebral Cortex*, *1*, 62–79.

Dias, R., Robbins, T. W., & Roberts, A. C. (1997). Dissociable forms of inhibitory control within prefrontal cortex with an analog of the Wisconsin Card Sort Test: Restriction to novel situations and independence from "on-line" processing. *Journal of Neuroscience*, *17*, 9285–9297.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, *16*, 143–149.

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations and the hippocampus: II. A computational analysis. *Hippocampus*, *13*, 341–54.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 875–894.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl type card sorting problem. *Journal of Experimental Psychology*, *38*, 404–411.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin card sorting test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources Inc.

Herd, S. A., Banich, M. T., & O'Reilly, R. C. (submitted). Neural mechanisms of cognitive control: An integrative model of stroop task performance and fMRI data.

Kroger, J., Nystrom, L., O'Reilly, R. C., Noelle, D., Braver, T. S., & Cohen, J. D. (in preparation). Multiple levels of temporal abstraction in the prefronal cortex: converging results from a computational model and fmri.

Levine, D. S., & Prueitt, P. S. (1989). Modeling some effects of frontal lobe damage-novelty and perseveration. *Neural Networks*, *2*, 103–116.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing. volume 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Miller, E. K., & Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science*, *263*, 520–522.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefontal cortex of the macaque. *Journal of Neuroscience*, *16*, 5154.

Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology*, *9*, 90–100.

Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: A neural network model of perseveration and dissociation in early childhood. *Developmental Psychobiology*, *40*, 255–265.

Munakata, Y., & Yerys, B. E. (2001). All together now: When dissociations between knowledge and action disappear. *Psychological Science*, *12*, 335–337.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*, 611–646.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199–1242.

O'Reilly, R. C. (in press). The division of labor between the neocortex and hippocampus. In G. Houghton (Ed.), *Connectionist modeling in cognitive psychology*. Psychology Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Noelle, D., Braver, T. S., & Cohen, J. D. (2002). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control. *Cerebral Cortex*, *12*, 246–257.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *6*, 505–510.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.

O'Reilly, R. C., & Soto, R. (2002). A model of the phonological loop: Generalization and binding. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.

Owen, A. M., Roberts, A. C., Hodges, J. R., Summers, B. A., Polkey, C. E., & Robbins, T. W. (1993). Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain*, *116*, 1159–1175.

Paulesu, E., Frith, C. D., & Frackowiak, R. S. J. (1993). The neural correlates of the verbal component of working memory. *Nature*, *362*, 342–345.

Roberts, A. C., Robbins, T. W., & Everitt, B. J. (1988). The effects of intradimensional and extradimensional shifts on visual discrimination learning in humans and non-human primates. *Quarterly Journal of Experimental Psychology*, *40*, 321–341.

Rogers, R. D., Andrews, T. C., Grasby, P. M., Brooks, D. J., & Robbins, T. W. (2000). Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, *12*, 142–162.

Rougier, N. P., & O'Reilly, R. C. (2002). Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*, *26*, 503–520.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1 and 2. Cambridge, MA: MIT Press.

Servan-Schreiber, D., Cohen, J. D., & Steingard, S. (1997). Schizophrenic deficits in the processing of context: A test of a theoretical model. *Archives of General Psychiatry*, *53*, 1105–1113.

Shallice, T., & Vallar, G. (1990). The impairment of auditory-verbal short-term storage. In G. Vallar, & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 11–53). Cambridge, England: Cambridge University Press.

Smith, E. E., & Jonides, J. (1998). Neuroimaging analyses of human working memory. *Proceedings of the National Academy of Sciences*, *95*, 12061.

Sternberg, S. (1966). High speed scanning in human menory. *Science*, *153*, 652–654.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, *11*, 37–63.