Cognitive, Affective, & Behavioral Neuroscience
2001, 1 (2), 137-160

Interactions between frontal cortex and basal ganglia in working memory: A computational model

MICHAEL J. FRANK, BRYAN LOUGHRY, and RANDALL C. O'REILLY
University of Colorado, Boulder, Colorado

The frontal cortex and the basal ganglia interact via a relatively well understood and elaborate system of interconnections. In the context of motor function, these interconnections can be understood as disinhibiting, or "releasing the brakes," on frontal motor action plans: The basal ganglia detect appropriate contexts for performing motor actions and enable the frontal cortex to execute such actions at the appropriate time. We build on this idea in the domain of working memory through the use of computational neural network models of this circuit. In our model, the frontal cortex exhibits robust active maintenance, whereas the basal ganglia contribute a selective, dynamic gating function that enables frontal memory representations to be rapidly updated in a task-relevant manner. We apply the model to a novel version of the continuous performance task that requires subroutine-like selective working memory updating and compare and contrast our model with other existing models and theories of frontal-cortex–basal-ganglia interactions.

It is almost universally accepted that the prefrontal cortex (PFC) plays a critical role in working memory, even though there is little agreement about exactly what working memory is or how *else* the prefrontal cortex contributes to cognition. Furthermore, it has long been known that the basal ganglia interact closely with the frontal cortex (e.g., Alexander, DeLong, & Strick, 1986) and that damage to the basal ganglia can produce many of the same cognitive impairments as damage to the frontal cortex (e.g., L. L. Brown, Schneider, & Lidsky, 1997; R. G. Brown & Marsden, 1990; Middleton & Strick, 2000b). This close relationship raises many questions regarding the cognitive role of the basal ganglia and how it can be differentiated from that of the frontal cortex itself. Are the basal ganglia and frontal cortex just two undifferentiated pieces of a larger system? Do the basal ganglia and the frontal cortex perform essentially the same function but operate on different domains of information/processing? Are the basal ganglia an evolutionary predecessor to the frontal cortex, with the frontal cortex performing a more sophisticated version of the same function?

We attempt to answer these kinds of questions by presenting a mechanistic theory and an implemented computational model of the contributions of the prefrontal cortex and basal ganglia to working memory. We find that the somewhat Byzantine nature of the anatomical loops connecting the frontal cortex and the basal ganglia make good computational sense in terms of a well-defined characterization of working memory function. Specifically, we argue that working memory requires *rapid updating* and *robust maintenance* as achieved by a *selective gating mechanism* (Braver & Cohen, 2000; Cohen, Braver, & O'Reilly, 1996; O'Reilly, Braver, & Cohen, 1999; O'Reilly & Munakata, 2000). Furthermore, although the frontal cortex and the basal ganglia are mutually interdependent in our model, we can nevertheless provide a precise division of labor between these systems. On this basis, we can make a number of specific predictions regarding the differential effects of frontal versus basal ganglia damage on a variety of cognitive tasks.

We begin with a brief overview of working memory, highlighting what we believe are the critical functional demands of working memory that the biological substrates of the frontal cortex and basal ganglia must subserve. We show that these functional demands can be met by a selective gating mechanism, which can trigger the updating of some elements in working memory while others are robustly maintained. Building on existing, biologically based ideas about the role of the basal ganglia in working memory (e.g., Beiser & Houk, 1998; Dominey, 1995), we show that the basal ganglia are well suited for providing this selective gating mechanism. We then present a neural network model that instantiates our ideas and performs a working memory task that requires a selective gating mechanism. We also show that this network can account for the role of the basal ganglia in sequencing tasks.

Versions of this material provided partial fulfillment of master's theses for M.J.F. and B.L. This work was supported by NIH Program Project Grant MH47566, NSF Grant IBN-9873492, and ONR Grant N00014-00-1-0246. We thank David Noelle, Todd Braver, Clay Holroyd, Jonathan Cohen, Seth Herd, and Marie Banich for helpful comments and discussion. Correspondence concerning this article should be addressed to R. C. O'Reilly, Department of Psychology, University of Colorado, 345 UCB, Boulder, CO 80309 (e-mail: oreilly@psych.colorado.edu).

We conclude by discussing the relationship between this model and other existing models of the basal-ganglia–frontal-cortex system.

WORKING MEMORY

Working memory can be defined as an active system for temporarily storing and manipulating information needed for the execution of complex cognitive tasks (Baddeley, 1986). For example, this kind of memory is clearly important for performing mental arithmetic (e.g., multiplying 42 × 17)—one must maintain subsets of the problem (e.g., 7 × 2) and store partial products (e.g., 14) while maintaining the original problem as well (e.g., 42 and 17; see e.g., Tsung & Cottrell, 1993). It is also useful in problem solving (maintaining and updating goals and subgoals, imagined consequences of actions, etc.), language comprehension (keeping track of many levels of discourse, using prior interpretations to correctly interpret subsequent passages, etc.), and many other cognitive activities (see Miyake & Shah, 1999, for a recent survey).

From a neural perspective, one can identify working memory with the maintenance and updating of information encoded in the active firing of neurons (*activation-based memory*; see e.g., Fuster, 1989; Goldman-Rakic, 1987). It has long been known that the PFC exhibits this kind of sustained active firing over delays (e.g., Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971; Kubota & Niki, 1971; Miller, Erickson, & Desimone, 1996; Miyashita & Chang, 1988). Such findings support the idea that the PFC is important for active maintenance of information in working memory.

The properties of this activation-based memory can be understood by contrasting them with more long-term kinds of memories that are stored in the synaptic connections between neurons (*weight-based memory*; Cohen et al., 1996; Munakata, 1998; O'Reilly et al., 1999; O'Reilly & Munakata, 2000). Activation-based memories have a number of advantages, relative to weight-based memories. For example, activation-based memories can be rapidly updated just by changing the activation state of a set of neurons. In contrast, changing weights requires structural changes in neural connectivity, which can be much slower. Also, information maintained in an active state is directly accessible to other parts of the brain (i.e., as a constant propagation of activation signals to all connected neurons), whereas synaptic changes only directly affect the neuron on the receiving end of the connection, and then only when the sending neuron is activated. In more familiar terms, activation-based memories are like sending a message via broadcast radio signal, whereas weight-based memories are like sending a letter in the mail.

These mechanistic properties of activation-based memories coincide well with oft-discussed characteristics of information maintained in working memory. Specifically, working memory is used for processing because it can be rapidly updated to reflect the ongoing products and demands of processing, and it is generally consciously accessible and can be described in a verbal protocol (e.g., Miyake & Shah, 1999). Furthermore, the active nature of working memory provides a natural mechanism for *cognitive control* (also known as *task-based attention*), where top-down activation can influence processing elsewhere to achieve task-relevant objectives (Cohen, Dunbar, & McClelland, 1990; Cohen & O'Reilly, 1996; O'Reilly et al., 1999). Thus, working memory and cognitive control can be seen as two different manifestations of the same underlying mechanism of actively maintained information. Of course, these manifestations function in different ways in different tasks and, thus, are not the same *psychological* construct, but both can be subserved by a common mechanism.

However, with these advantages of activation-based memories there are also concomitant disadvantages. For example, because these memories do not involve structural changes, they are transient and, therefore, do not provide a suitable basis for long-term memories. Also, because information is encoded by the activation states of neurons, the capacity of these memories scales as a function of the number of neurons, whereas the capacity of weight-based memories scales as a function of the number of synaptic connections, which is much larger.

Because of this fundamental tradeoff between activation- and weight-based memory mechanisms, it makes sense that the brain would have evolved two different specialized systems to obtain the best of both types of memory. This is particularly true if there are specific mechanistic specializations that are needed to make each type of memory work better. There has been considerable discussion along these lines of the ways in which the neural structure of the hippocampus is optimized for subserving a particular kind of weight-based memory (e.g., McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2000, 2001). Similarly, this paper represents the further development of a line of thinking about the ways in which the frontal cortex is specialized to subserve activation-based memory (Braver & Cohen, 2000; Cohen et al., 1996; O'Reilly et al., 1999). In the next section, we will introduce a specific working memory task that exemplifies the functional specializations needed to support effective activation-based memories, and we then proceed to explore how the biology of the frontal-cortex–basal-ganglia system is specialized to achieve these functions.

Working Memory Functional Demands

The A–X version of the continuous performance task (CPT–AX) is a standard working memory task that has been extensively studied in humans (Braver & Cohen, 2000; Cohen et al., 1997). The subject is presented with sequential letter stimuli (A, X, B, Y) and is asked to detect the specific sequence of an A followed by an X by pushing the right button. All other combinations (A–Y, B–X, B–Y) should be responded to with a left button push. This task requires a relatively simple form of working memory, where the prior stimulus must be maintained over a delay until the next stimulus appears, so that one can discriminate the target from the nontarget sequences. We have

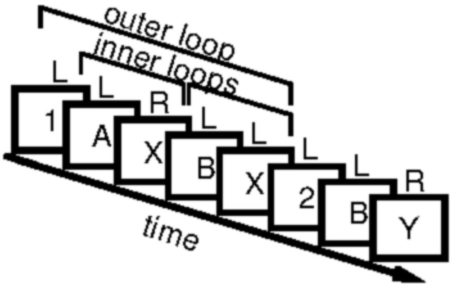

Figure 1. The 1–2–AX version of the continuous performance task. Stimuli are presented one at a time in a sequence, and the subject must respond by pressing the right key (R) to the target sequence; otherwise, a left key is pressed. If the subject last saw a 1, the target sequence is an A followed by an X. If a 2 was last seen, then the target is a B followed by a Y. Distractor stimuli (e.g., 3, C, Z) may be presented at any point in a sequence and are to be ignored. Shown is an example sequence of stimuli and the correct responses, emphasizing the inner- and outer-loop nature of the memory demands (maintaining the task stimuli [1 or 2] is an outer loop, whereas maintaining the prior stimulus of a sequence is an inner loop).

devised an extension of this task that places somewhat more demands on the working memory system. In this extension, which we call the 1–2–AX task (Figure 1), the target sequence varies depending on prior *task demand* stimuli (a 1 or a 2). Specifically, if the subject last saw a 1, the target sequence is A–X. However, if the subject last saw a 2, the target sequence is B–Y.[1] Thus, the task demand stimuli define an *outer loop* of active maintenance (maintenance of task demands), within which there can be a number of *inner loops* of active maintenance for the A–X level sequences.

The full 1–2–AX task places three critical functional demands on the working memory system.

1. *Rapid updating.* As each stimulus comes in, it must be rapidly encoded in working memory (e.g., one-trial updating, which is not easily achieved in weight-based memory).

2. *Robust maintenance.* The task demand stimuli (1 or 2) in the outer loop must be maintained in the face of interference from ongoing processing of inner loop stimuli and irrelevant distractors.

3. *Selective updating.* Only some elements of working memory should be updated at any given time, while others are maintained. For example, in the inner loop, As and Xs (etc.) should be updated while the task demand stimulus (1 or 2) is maintained.

One can obtain some important theoretical leverage by noting that the first two of these functional demands are directly in conflict with each other, when viewed in terms of standard neural processing mechanisms (Braver & Cohen, 2000; Cohen et al., 1996; O'Reilly et al., 1999; O'Reilly & Munakata, 2000). Specifically, rapid updating can be achieved by making the connections between stimulus input and working memory representations strong, but this directly impairs robust maintenance, since such strong connections would allow stimuli to interfere with ongoing maintenance. This conflict can be resolved by using an active *gating* mechanism (Cohen et al., 1996; Hochreiter & Schmidhuber, 1997).

Gating

An active gating mechanism dynamically regulates the influence of incoming stimuli on the working memory system (Figure 2). When the gate is open, stimulus information is allowed to flow strongly into the working memory system, thereby achieving rapid updating. When the gate is closed, stimulus information does not strongly influence working memory, thereby allowing robust maintenance in the face of ongoing processing. The computational power of such a gating mechanism has been demonstrated in the LSTM model of Hochreiter and Schmidhuber (1997), which is based on error backpropagation mechanisms and has not been related to brain function, and in more biologically based models by Braver and Cohen (2000) and O'Reilly and Munakata (2000).

These existing biologically based models provide the point of departure for the present model. These models were based on the idea that the neuromodulator *dopamine* can perform the gating function by transiently strengthening the efficacy of other cortical inputs to the frontal cortex. Thus, when dopamine release is phasically elevated, as has been shown in a number of neural recordings (e.g., Schultz, Apicella, & Ljungberg, 1993), working memory can be updated. Furthermore, these models incorporate the intriguing idea that the same factors that drive dopamine spikes for learning (e.g., Montague, Dayan, & Sejnowski, 1996) should also be appropriate for driving working memory updating. Specifically, working memory should be updated whenever a stimulus triggers an enhanced prediction of future reward. However, an important limitation of these models comes from the fact that dopamine release is relatively global; large areas

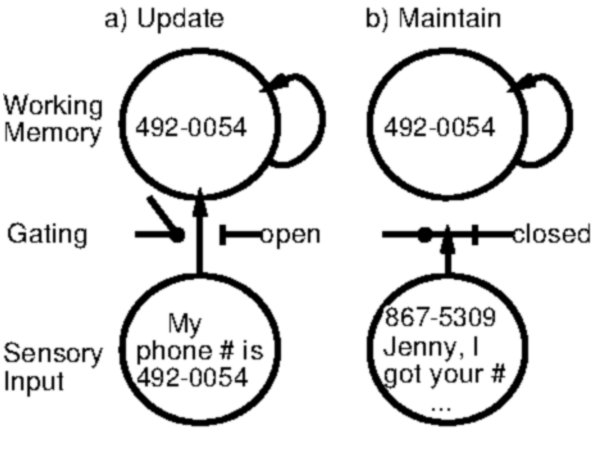

Figure 2. Illustration of active gating. When the gate is open, sensory input can rapidly update working memory (e.g., allowing one to store a phone number), but when it is closed, it cannot, thereby preventing other distracting information (e.g., an irrelevant phone number) from interfering with the maintenance of previously stored information.

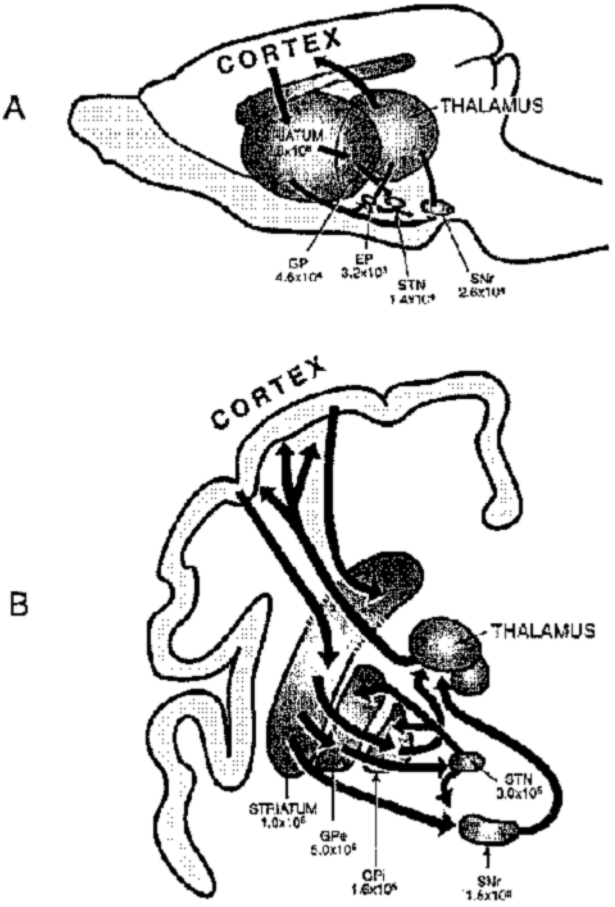

Figure 3. Schematic diagram of the major structures of the basal ganglia and their connectivity with the frontal cortex in the rat (A) and human (B). GP, globus pallidus; GPi, GP internal segment; GPe, GP external segment; SNr, substantia nigra pars reticulata; EP, entopeduncular nucleus; STN, subthalamic nucleus. Numbers indicate total numbers of neurons within each structure. Note that the EP in rodents is generally considered homologous to the GPi in primates, and the GP in rodents is homologous to the GPe in primates. From "Basal Ganglia: Structure and Computations," by J. Wickens, 1997, *Network: Computation in Neural Systems*, *8*, p. R79. Copyright 1997 by IOP Publishing Ltd. Reprinted with permission.

of the PFC would therefore receive the same gating signal. In short, a dopamine-based gating mechanism does not support the *selective* updating functional demand listed above, where some working memory representations are updated as others are being robustly maintained. Therefore, the present model explores the possibility that the basal ganglia can provide this selective gating mechanism, as will be described next.

THE BASAL GANGLIA AS A SELECTIVE GATING MECHANISM

Our model is based directly on a few critical features of the basal-ganglia–frontal-cortex system, which we review here. Figures 3 and 4 show schematic diagrams of the relevant circuitry. At the largest scale, one can see a number of *parallel loops* from the frontal cortex to the striatum (also called the neostriatum, consisting of the caudate nucleus, putamen, and nucleus accumbens) to the globus pallidus internal segment (GPi) or substantia nigra pars reticulata (SNr) and then on to the thalamus, finally projecting back up in the frontal cortex (Alexander et al., 1986). The GPi and SNr circuits are functionally analogous (although they have different subcortical targets), so we consider them as one functional entity. Both the frontal cortex and the striatum also receive inputs from various areas of the posterior/sensory cortex. There are also other pathways within the basal ganglia involving the external segment of the globus pallidus and the subthalamic nucleus that we see as having a role in learning but are not required for the basic gating operation of the network; these other circuits only project through the GPi/SNr to affect frontal function.

The critical aspect of this circuit for gating is that the striatal projections to the GPi/SNr and from the GPi/SNr to the thalamus are *inhibitory.* Furthermore, the GPi/SNr neurons are *tonically active*, meaning that in the absence of any other activity, the thalamic neurons are inhibited by constant firing of GPi/SNr neurons. Therefore, when the striatal neurons fire, they serve to *disinhibit* the thalamic neurons (Chevalier & Deniau, 1990; Deniau & Chevalier, 1985). As was emphasized by Chevalier and Deniau (and was suggested earlier by others; Neafsey, Hull, & Buchwald, 1978; Schneider, 1987), this disinhibition produces a *gating* function (this is literally the term they used): It *enables* other functions to take place but does not directly *cause* them to occur, as a direct excitatory connection would. Chevalier and Deniau review a range of findings from the motor control domain, showing that the activation of striatal neurons enables, but does not directly cause, subsequent motor movements.

In short, one can think of the overall influence of the basal ganglia on the frontal cortex as "releasing the brakes" for motor actions and other functions. Put another way, the basal ganglia are important for *initiating* motor movements, but not for determining the detailed properties of these movements (e.g., Bullock & Grossberg, 1988; Chevalier & Deniau, 1990; Hikosaka, 1989; Passingham, 1993). Clearly, this disinhibitory gating in the motor domain could easily be extended to gating in the working memory domain. Indeed, this suggestion was made by Chevalier and Deniau in generalizing their ideas from the motor domain to the cognitive one. Subsequently, several theories and computational models have included variations of this idea (Alexander, Crutcher, & DeLong, 1990; Beiser & Houk, 1998; Dominey, 1995; Dominey & Arbib, 1992; Gelfand, Gullapalli, Johnson, Raye, & Henderson, 1997; Goldman-Rakic & Friedman, 1991; Houk & Wise, 1995). Thus, we find a striking convergence between the functionally motivated gating ideas we presented earlier and similar ideas developed more from a bottom-up consideration of the biological properties of the basal-ganglia–frontal-cortex system.

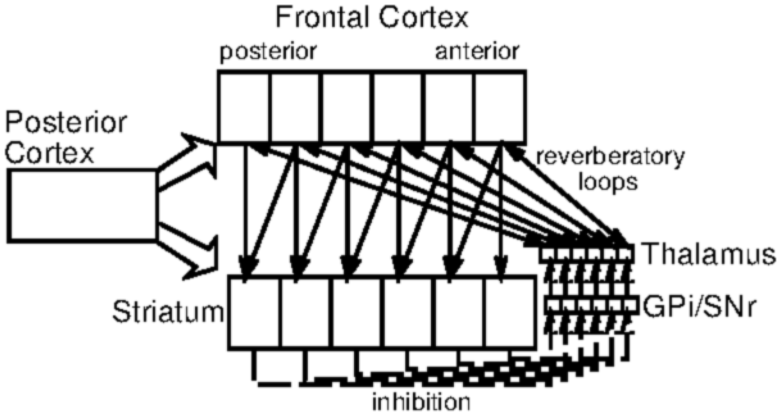

Figure 4. The basal ganglia (striatum, globus pallidus, and thalamus) are interconnected with the frontal cortex through a series of parallel loops. Excitatory connections are in solid lines, and inhibitory ones are in dashed lines. The frontal cortex projects excitatory connections to the striatum, which then projects inhibition to the globus pallidus internal segment (GPi) or the substantia nigra pars reticulata (SNr), which again project inhibition to nuclei in the thalamus, which are reciprocally interconnected with the frontal cortex. Because GPi/SNr neurons are tonically active, they are constantly inhibiting the thalamus, except when the striatum fires and disinhibits the thalamus. This disinhibition provides a modulatory or gating-like function.

Specifically, in the context of the working memory functions of the frontal cortex, our model is based on the idea that the basal ganglia are important for *initiating the storage of new memories.* In other words, the disinhibition of the thalamocortical loops by the basal ganglia results in the opening of the gate into working memory, resulting in rapid updating. In the absence of striatal firing, this gate remains closed, and the frontal cortex maintains existing information. Critically, the basal ganglia can provide a *selective* gating mechanism because of the many parallel loops. Although the original neuroanatomical studies suggested that there are around five such loops (Alexander et al., 1986), it is likely that the anatomy can support many more subloops within these larger scale loops (e.g., Beiser & Houk, 1998), meaning that relatively fine-grained selective control of working memory is possible. We will discuss this in greater detail later.

To summarize, at least at this general level, it appears that the basal ganglia can provide exactly the kind of selective gating mechanism that our functional analysis of working memory requires. Our detailed hypotheses regarding the selective gating mechanisms of this system are specified in the following sections.

Details of Active Maintenance and the Gating Mechanism

We begin with a discussion of the mechanisms of active maintenance in the frontal cortex, which then constrain the operation of the gating mechanism provided by the basal ganglia.

Perhaps the most obvious means of achieving the kinds of actively maintained neural firing observed in PFC neurons using basic neural mechanisms is to have *recurrent excitation* among frontal neurons, resulting in *attractor states* that persist over time (e.g., Braver & Cohen, 2000; Dehaene & Changeux, 1989; Moody, Wise, di Pellegrino, & Zipser, 1998; O'Reilly & Munakata, 2000; Seung, 1998; Zipser, Kehoe, Littlewort, & Fuster, 1993). With this kind of mechanism, active maintenance is achieved because active neurons will provide further activation to themselves, perpetuating an activity state. Most of the extant theories/models of the basal ganglia role in working memory employ a variation of this type of maintenance, where the recurrent connections are between frontal neurons and the thalamus and back (Alexander et al., 1990; Beiser & Houk, 1998; Dominey, 1995; Dominey & Arbib, 1992; Gelfand et al., 1997; Goldman-Rakic & Friedman, 1991; Hikosaka, 1989; Houk &Wise, 1995; Taylor & Taylor, 2000). This form of recurrence is particularly convenient for enabling the basal ganglia to regulate the working memory circuits, since thalamic disinhibition would directly facilitate the flow of excitation through the thalamocortical loops.

However, it is unclear whether there are sufficient numbers of thalamic neurons, relative to frontal neurons, to support the full space of maintainable frontal representations. When a given thalamic neuron sends activation to the frontal cortex to support maintenance, its connectivity would have to uniquely support one particular representation, or part thereof; otherwise, the specificity of the maintained information would be lost. Therefore, the number of thalamic neurons would have to be on the same order as that of the frontal neurons, unless frontal representations are massively redundant. Recurrent connectivity within the frontal cortex itself avoids this problem. Furthermore, we are not aware of any definitive evidence suggesting that these loops are indeed critical for active maintenance (e.g., showing that frontal active maintenance is eliminated with selective thalamic lesions, which is presumably a feasible experiment). Another issue with thalamocortically mediated recurrent loops is that they would generally require persistent disinhibition in the thalamus during the entire maintenance period (although see Beiser & Houk, 1998, for a way of avoiding this constraint). For these reasons, we are inclined to think in terms of intracortical recurrent connectivity for supporting frontal maintenance.

Although it is intuitively appealing, the recurrence-based mechanism has some important limitations stemming from the fact that information maintenance is entirely dependent on the instantaneous activation state of the network. For example, it does not allow for the frontal cortex's exhibiting a transient, stimulus-driven activation state and then returning to maintaining some previously encoded information—the set of neurons that are most active at any given point in time will receive the strongest excitatory recurrent feedback and will, therefore, be what is maintained. If a transient stimulus activates frontal neurons above the level of previously maintained information, this stimulus transient will displace the prior information as what is maintained.

This survival-of-the-most-active characteristic is often violated in recordings of prefrontal cortex neurons. For example, Miller et al. (1996) observed that frontal neurons will tend to be activated transiently when irrelevant stimuli are presented while monkeys are maintaining other task-relevant stimuli. During these stimulus transients, the neural firing for the maintained stimulus can be weaker than that for the irrelevant stimulus. After the irrelevant stimuli disappear, the frontal activation reverts to maintaining the task-relevant stimuli. We interpret this data as strongly suggesting that frontal neurons have some kind of *intrinsic* maintenance capabilities.[2] This means that individual frontal neurons have some kind of intracellular "switch" that, when activated, provides these neurons with extra excitatory input that enhances their capacity to maintain signals in the absence of external input. Thus, this extra excitation enables maintaining neurons to recover their activation state after a stimulus transient: After the actual stimulus ceases to support its frontal representation, the neurons with intrinsic excitation will dominate.

There are a number of possible mechanisms that could support a switchable intrinsic maintenance capacity for frontal neurons (e.g., Dilmore, Gutkin, & Ermentrout, 1999; Durstewitz, Seamans, & Sejnowski, 2000b; Fellous, Wang, & Lisman, 1998; Gorelova & Yang, 2000; Lewis & O'Donnell, 2000; Wang, 1999). For example, Lewis and O'Donnell report clear evidence that, at least in an anesthetized preparation, prefrontal neurons exhibit bistability—they have *up* and *down* states. In the *up* state, neurons have a higher resting potential and can easily fire spikes. In the *down* state, the resting potential is more negative, and it is more difficult to fire spikes. A number of different possible mechanisms are discussed by Lewis and O'Donnell that can produce these effects, including selective activation of excitatory ion channels in the *up* state (e.g., Ca^{2+} or Na^{+}) or selective activation of inhibitory K^{+} ion channels in the *down* state.

Other mechanisms that involve intracellular switching but depend more on synaptic input have also been proposed. These mechanisms take advantage of the properties of the NMDA receptor, which is activated both by synaptic input and by postsynaptic neuron depolarization and produces excitation through Ca^{2+} ions (Durstewitz, Kelc, & Gunturkun, 1999; Durstewitz, Seamans, & Sejnowski, 2000a; Fellous et al., 1998; Wang, 1999). In the model by Wang and colleagues, a switchable bistability emerges as a result of interactions between NMDA channels and the balance of excitatory and inhibitory inputs. In the model by Durstewitz and colleagues, dopamine modulates NMDA channels and inhibition to stabilize a set of active neurons and prevent interference from other neurons (via the inhibition). Consistent with these models, we think that recurrent excitation plays an important maintenance role, in addition to a switchable intrinsic maintenance capacity. As we will discuss below, recurrent excitation can provide a "default" maintenance function, and it is also important for magnifying and sustaining the effects of the intrinsic maintenance currents.

There are many complexities and unresolved issues with these maintenance mechanisms. For example, although dopamine clearly plays an important role in some of these mechanisms, it is not clear whether tonic levels present in awake animals would be sufficient to enable these mechanisms or whether phasic bursts of dopamine would be required. This can have implications for the gating mechanism, as we will discuss later. Despite the tentative nature of the empirical evidence, there are enough computational advantages to a switchable intrinsic maintenance capacity (as combined with a more conventional form of recurrent excitation) to compel us to use such a mechanism in our model. Furthermore, we think the neurophysiological finding that working memory neurons recover their memory-based firing even after representing transient stimuli (as was reviewed above) makes a compelling empirical case for the presence of such mechanisms.

There are two primary computational advantages to a switchable intrinsic maintenance capacity. The first is that it imparts a significant degree of robustness on active maintenance, as has been documented in several models (e.g., Durstewitz et al., 2000a; Fellous et al., 1998). This robustness stems from the fact that intrinsic signals are not dependent on network dynamics, whereas spurious strong activations can hijack recurrent maintenance mechanisms. Second, these intrinsic maintenance mechanisms, by allowing the frontal cortex to represent both transient stimuli and maintained stimuli, avoid an important catch-22 problem that arises in bootstrapping learning over delays (O'Reilly & Munakata, 2000; Figure 5). Briefly, learning that it is useful to maintain a stimulus can occur only after that stimulus has been maintained in frontal rep-

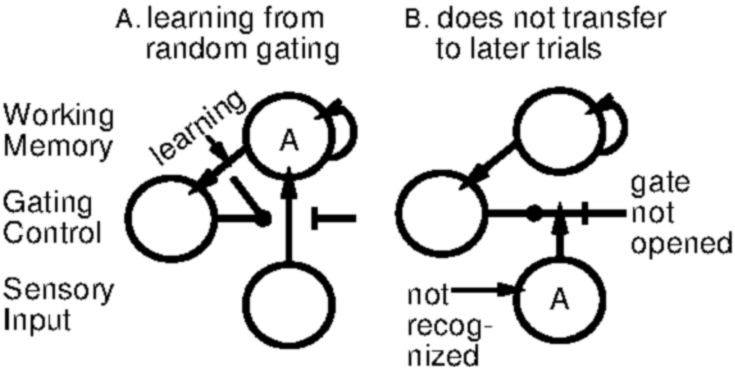

Figure 5. Illustration of the catch-22 problem that occurs when the gating mechanism learns on the basis of maintained working memory representations and those representations can become activated only after the gating mechanism fires for a given stimulus. (A) Learning about a stimulus A presented earlier and maintained in the frontal cortex, which is based on initially random exploratory gating signals, will be between the maintained representations and the gating controller. (B) When this stimulus is later presented, it will not activate the working memory representations until the gate is opened, but the gate has only learned about this stimulus from the same working memory representations, which are not activated.

resentations, meaning that the gating mechanism must learn what to maintain on the basis of frontal representations. However, if these frontal representations only reflect stimuli that have already been gated in for maintenance, the gating mechanism will not be able to detect this stimulus as something to gate in until it is already gated into the frontal cortex! However, if the frontal representations always reflect current stimuli as well as maintained information, this problem does not occur.

Dynamic gating in the context of an intracellular maintenance switch mechanism amounts to the activation and deactivation of this switch. Neurons that participate in the maintenance should have the switch turned on, and those that do not should have the switch turned off. This contrasts with other gating models developed in the context of recurrent activation-based maintenance, which required gating to modulate the strength of input weights into the frontal cortex (e.g., Braver & Cohen, 2000; O'Reilly & Munakata, 2000), or the strength of the thalamocortical recurrent loops (e.g., Beiser & Houk, 1998; Dominey, 1995; Gelfand et al., 1997). Therefore, we propose that the disinhibition of the thalamocortical loops by the basal ganglia results in the modulation of the intracellular switch. Specifically, we suggest that the activation of the Layer 4 frontal neurons that receive the excitatory projection from the thalamus (or the equivalent cell types in Layer 3 in motor areas of the frontal cortex—we will just use the Layer 4 notation, for convenience) is responsible for modulating intracellular ion channels on the neurons in other layers (which could be in either Layers 2–3 or 5–6) that are ultimately responsible for maintaining the working memory representations.

In our model, we further specify that the intracellular switch is activated when a neuron is receiving strong excitatory input from other areas (e.g., stimulus input) in addition to the Layer 4 input, and it is deactivated if the Layer 4 input does not coincide with other strong excitatory input. Otherwise, the switch just stays in its previous state (and is, by default, off). This mechanism works well in practice for appropriately updating working memory representations and could be implemented through the operation of NMDA channels that require a conjunction of postsynaptic depolarization and synaptic input (neurotransmitter release). Alternatively, such NMDA channels could also activate other excitatory ion channels via second messengers, or other voltage-gated channels could directly mediate the effect, so we are at present unsure as to the exact biological mechanisms necessary to implement such a rule. Nevertheless, the overall behavior of the ion channels is well specified and could be tested with appropriate experiments.

Finally, more conventional recurrent excitation-based maintenance is important in our model for establishing a "default" propensity of the frontal cortex to maintain information. Thus, if nothing else has been specifically gated on in a region of the frontal cortex (i.e., if no other neurons have a specific competitive advantage owing to intracellular maintenance currents), the recurrent connectivity will tend to maintain representations over time anyway. However, any new stimulus information will easily displace this kind of maintained information, and it cannot compete with information that has been specifically gated on. This default maintenance capacity is important for "speculative" trial-and-error maintenance of information; the only way for a learning mechanism to discover whether it is important to maintain something is if it actually does maintain it and, then, it turns out to be important. Therefore, having a default bias to maintain is useful. However, this default maintenance bias is overridden by the active gating mechanism, allowing learning to have full control over what is ultimately maintained.

To summarize, in our model, active maintenance operates according to the following set of principles.

1. Stimuli generally activate their corresponding frontal representations when they are presented.

2. Robust maintenance occurs only for those stimuli that trigger the intracellular maintenance switch (as a result of the conjunction of external excitation from other cortical areas and Layer 4 activation resulting from basal-ganglia-mediated disinhibition of the thalamocortical loops).

3. When other stimuli are being maintained, those representations that did not have the intracellular switch activated will decay quickly following stimulus offset.

4. However, if nothing else is being maintained, recurrent excitation is sufficient to maintain a stimulus until other stimuli are presented. This "default" maintenance is important for learning, by trial and error, what it is relevant to maintain.

Additional Anatomical Constraints

In this section, we will discuss the implications of a few important anatomical properties of the basal-ganglia–frontal-cortex system. First, we will consider the consequences of the relative sizes of different regions in the basal-ganglia frontal cortex pathway. Next, we will examine evidence that can inform the number of different, separately gatable frontal areas. Finally, we will discuss the level of convergence and divergence of the loops.

A strong constraint on understanding basal ganglia function comes from the fact that the GPi and SNr have a relatively small number of neurons—there are approximately 111 million neurons in the human striatum (Fox & Rafols, 1976), whereas there are only 160,000 in the GPi (Lange, Thorner, & Hopf, 1976) and a similar number in the SNr. This means that whatever information is encoded by striatal neurons must be vastly compressed or eliminated on its way up to the frontal cortex. This constraint coincides nicely with the gating hypothesis: The basal ganglia do not need to convey detailed *content* information to the frontal cortex; instead, they simply need to tell different regions of the frontal cortex *when* to update. As we noted in the context of motor control, damage to the basal ganglia appears to affect *initiation*, but not the details of *execution*, of motor movements—presumably, not that many neurons are needed to encode this gating or initiation information.

Given this dramatic bottleneck in the GPi/SNr, one might wonder why there are so many striatal neurons in the first place. We think this is also sensible under the gating proposal: In order for only task-relevant stimuli to get updated (or an action initiated) via striatal firing, these neurons need to fire only for a very specific *conjunction* of environmental stimuli and internal context representations (as conveyed through descending projections from the frontal cortex). This context specificity of striatal firing has been established empirically (e.g., Schultz, Apicella, Romo, & Scarnati, 1995) and is an important part of many extant theories/models (e.g., Amos, 2000; Beiser & Houk, 1998; Berns & Sejnowski, 1996; Houk & Wise, 1995; Jackson & Houghton, 1995; Wickens, 1993; Wickens, Kotter, & Alexander, 1995). Thus, many striatal neurons are required to encode all of the different specific conjunctions that can be relevant. Without such conjunctive specificity, there would be a risk that striatal neurons would fire for inappropriate subsets of stimuli. For example, the 1 and 2 stimuli should be maintained separately from the other stimuli in the 1–2–AX task, but this is not likely to be true of other tasks. Therefore, striatal neurons should encode the conjunction of the stimulus (1 or 2) together with some representation of the 1–2–AX task context from the frontal cortex. If the striatum instead employed a smaller number of neurons that just responded to stimuli without regard to task context (or other similar kinds of conjunctions), confusions between the many different implications of a given stimulus would result. Note that by focusing on conjunctivity in the striatum, we do not mean to imply that there is no conjunctivity in the frontal representations as well (e.g., S. C. Rao, Rainer, & Miller, 1997; Watanabe, 1992); frontal conjunctive representations can be useful for maintaining appropriately contextualized information.

Another constraint to consider concerns the number of different subregions of the frontal cortex for which the basal ganglia can plausibly provide separate gating control. Although it is impossible to determine any precise estimates of this figure, even the very crude estimates we provide here are informative in suggesting that gating occurs at a relatively fine-grained level. Fine-grained gating is important for mitigating conflicts where two representations require separate gating control and yet fall within one gating region. An upper limit estimate is provided by the number of neurons in the GPi/SNr, which is roughly 320,000 in the human, as was noted previously. This suggests that the gating signal operates on a *region* of frontal neurons, instead of individually controlling specific neurons (and, assuming that the thalamic areas projecting to the frontal cortex are similarly sized, argues against the notion that the thalamocortical loops themselves can maintain detailed patterns of activity).

An interesting possible candidate for the regions of the frontal cortex that are independently controlled by the basal ganglia are distinctive anatomical structures consisting of interconnected groups of neurons, called *stripes* (Levitt, Lewis, Yoshioka, & Lund, 1993; Pucak, Levitt, Lund, & Lewis, 1996). Each stripe appears to be isolated from the immediately adjacent tissue but interconnected with other more distal stripes, forming a cluster of interconnected stripes. Furthermore, it appears that connectivity between the PFC and the thalamus exhibits a similar, although not identical, kind of discontinuous stripelike structuring (Erickson & Lewis, 2000). Therefore, it would be plausible that each stripe or cluster of stripes constitutes a separately controlled group of neurons;

each stripe can be separately updated by the basal ganglia system. Given that each stripe is roughly 0.2–0.4 × 2–4 mm in size (i.e., 0.4–1.6 mm^2 in area), one can make a rough computation that the human frontal cortex (having roughly one fourth of the approximately 140,000 mm^2 surface area of the entire cortex; Douglas & Martin, 1990) could have over 20,000 such stripes (assuming that the stripes found in monkeys also exist in humans, with similar properties). If the thalamic connectivity were with stripe clusters, and not individual stripes, this figure would be reduced by a factor of around five. In either case, given the size of the GPi and SNr, there would be some degree of redundancy in the per stripe gating signal at the GPi/SNr level. Also note that the 20,000 (or 4,000 for stripe clusters) figure is for the entire frontal cortex, so the proportion located in the PFC (and thus involved in working memory function) would be smaller. Further evidence consistent with the existence of such stripelike structures comes from the finding of isocoding microcolumns of neighboring neurons that all encode roughly the same information (e.g., having similar directional coding in a spatial delayed response task; S. G. Rao, Williams, & Goldman-Rakic, 1999).

The precise nature of the inputs and outputs of the loops through the basal ganglia can have implications for the operation of the gating mechanism. From a computational perspective, it would be useful to control each stripe by using a range of different input signals from the sensory and frontal cortex (i.e., broad convergence of inputs), to make the gating appropriately context specific. In addition, it is important to have input from the current state of the stripe that is being controlled, since this would affect whether this stripe should be updated or not. This implies closed loops going through the same frontal region. Data consistent with both of these connectivity patterns has been presented (see Graybiel & Kimura, 1995, and Middleton & Strick, 2000a, for reviews). Although some have taken mutually exclusive positions on these two patterns of connectivity and the facts are a matter of considerable debate, both patterns are mutually compatible from the perspective of our model. Furthermore, even if it turns out that the cortical projections to the striatum are relatively focused, context sensitivity in gating can be achieved via context-sensitive frontal input representations. In other words, the context sensitivity of gating could come either from focused context-sensitive inputs to the striatum or from broad sensory inputs that are integrated by the striatum itself. One particularly intriguing suggestion is that the convergence of inputs from other frontal areas may be arranged in a hierarchical fashion, providing a means for more anterior frontal areas (which may represent higher level, more abstract task/goal information) to appropriately contextualize more posterior areas (e.g., supplementary and primary motor areas; Gobbel, 1997). This hierarchical structure is reflected in Figure 4.

There are two aspects of the basal ganglia connectivity that we have not yet integrated into our model and thus stand as challenges for future work. First, the basal ganglia circuits through the thalamus also project to posterior cortex areas (the inferior temporal and parietal cortex) in addition to the frontal cortex (e.g., Middleton & Strick, 2000a). It is thus possible that gating occurs in these areas as well, but it is not clear that they are essential for robust working memory function (e.g., Constantinidis & Steinmetz, 1996; Miller et al., 1996). Therefore, we are not sure what functional role these connections play. Second, striatal neurons receive a substantial projection from the same thalamic areas that they disinhibit (e.g., McFarland & Haber, 2000). This projection has been largely ignored in computational and theoretical models of the basal ganglia but could have important implications. For example, such a projection would quickly inform striatal neurons about exactly which frontal regions were actually updated and could thus provide useful constraints on the allocation of subsequent updating to unused regions.

To summarize, anatomical constraints are consistent with the selective gating hypothesis by suggesting that the basal ganglia interacts with a large number of distinct regions of the frontal cortex. We hypothesize that these distinct stripe structures constitute separately gated collections of frontal neurons, extending the parallel loops concept of Alexander et al. (1986) to a much finer grained level (see also Beiser & Houk, 1998). Thus, it is possible to maintain some information in one set of stripes, while *selectively* updating other stripes.

Learning and the Role of Dopamine

Implicit in our gating model is that the basal ganglia somehow know when it is appropriate to update working memory representations. To avoid some kind of homunculus in our model, we posit that learning is essential for shaping the striatal firing in response to task demands. This dovetails nicely with the widely acknowledged role that the basal ganglia, and the neuromodulator dopamine, play in reinforcement learning (e.g., Barto, 1995; Houk, Adams, & Barto, 1995; Schultz, Dayan, & Montague, 1997; Schultz, Romo, et al., 1995). However, our work on integrating learning mechanisms with the basal ganglia selective gating model is still in progress. Therefore, our current model, presented in this paper, uses hand-wired representations (i.e., ones causing the striatum to fire only for task-relevant stimuli) to demonstrate the basic gating capacity of the overall system.

In addition to shaping striatal neurons to fire at the right time through stimulus-specific, phasic firing, dopamine may also play an important role in regulating the overall excitability of striatal neurons in a tonic manner. The gating model places strong demands on these excitability parameters, because striatal neurons need to be generally silent, while still being capable of firing when the appropriate stimulus and contextual inputs are present. This general silence, which is a well-known property of striatal neurons (e.g., Schultz, Apicella, et al., 1995) can be accomplished by having a relatively high effective threshold for firing (either because the threshold itself is high or because they experience more inhibitory currents that offset excitation). However, if this effective threshold is too high, striatal neurons will not be able to fire when the correct circumstances arise. Therefore, it is likely that the brain has developed specialized mechanisms for regulating these thresholds. The effects of Parkinson's disease, which results from a tonic loss of dopamine innervation of the basal ganglia, together with neurophysiological data showing dopaminergic modulation of different states of excitability in striatal neurons (e.g., Gobbel, 1995; Surmeier & Kitai, 1999; C. J. Wilson, 1993), all suggest that dopamine plays an important part in this regulatory mechanism.

The Motor-Control–Working-Memory Continuum

We have emphasized that our view of the basal ganglia interactions with the frontal cortex builds on existing ideas regarding these interactions in the context of motor control. Specifically, both the initiation of a motor act and the updating of working memory representations require striatal firing to disinhibit or gate frontal cortex representations. Although we have discussed motor control and working memory as two separable functions, it is probably more useful to think in terms of a continuum between cognitive working memory and motor control functions. For example, one can think of the neurons in premotor or supplementary motor areas as maintaining a motor control plan that guides a sequence of basic motor movements (e.g., Shima & Tanji, 1998; Wise, 1985). This plan would need to be maintained over the duration of the sequence and can thus be considered a working memory representation. Thus, the line between working memory and motor control is fuzzy; indeed, this ambiguity provides useful insight as to why both motor control and working memory are colocalized within the frontal cortex.

Summary: The Division of Labor Between Frontal Cortex and Basal Ganglia

Before describing our model in detail, and by way of summary, we return to the fundamental question posed at the outset of this paper: What is the nature of the division of labor between the frontal cortex and the basal ganglia? In light of all the foregoing information, we offer the following concise summary of this division of labor: The frontal cortex uses continuously firing activations to encode information over time in working memory (or, on a shorter time scale, to execute motor actions), and the basal ganglia fires only at very select times to trigger the updating of working memory states (or initiate motor actions) in the frontal cortex.

Furthermore, we can speculate as to *why* it would make sense for the brain to have developed this division of labor in the first place. Specifically, one can see that the use of continuously firing activation states to encode information is at odds with the need to fire only at very specific times. Therefore, the brain may have separated these two systems to develop specialized mechanisms supporting each. For example, striatal neurons must have a relatively high effective threshold for firing, and it can be difficult to regulate such a threshold to ensure that firing happens when appropriate, and not when it is not appropriate. The dopaminergic neuromodulation of these neurons and its control by descending projections from the striatum may be important specializations in this regard. Finally, we do not mean to claim that all striatal neurons fire only in a punctate manner; others exhibit sustained *delay period* activations (e.g., Alexander, 1987; Schultz, Romo, et al., 1995). We think that these reflect sustained frontal activations, not an intrinsic maintenance capability of striatal neurons themselves. Similarly, punctate firing in cortical neurons, especially in motor output areas, could be a reflection of gating signals from the basal ganglia.

THE 1–2–AX MODEL

We have implemented the ideas outlined above in a computational model of the 1–2–AX task. This model demonstrates how the basal ganglia can provide a selective gating mechanism, by showing that the outer-loop information of the task demand stimuli (1 or 2) can be robustly maintained while the inner-loop information (A, B, etc.) is rapidly updated. Furthermore, we show that irrelevant distractor stimuli are ignored by the model, even though they transiently activate their frontal representations. In addition, the model demonstrates that the same mechanisms that drive working memory updating also drive the motor responses in the model.

The Mechanics of the Model

The model is shown in Figure 6. The units in the model operate according to a simple *point neuron* function using rate-coded output activations, as implemented in the *Leabra* framework (O'Reilly, 1998; O'Reilly & Munakata, 2000). There are simulated excitatory synaptic input channels, and inhibitory input is computed through a simple approximation to the effects of inhibitory interneurons. There is also a constant leak current that represents the effects of K^+ channels that are always open, and the maintenance frontal neurons have a switchable excitatory ion channel that is off by default. (See the Appendix for the details and equations.) The model's representations were predetermined, but the specific weights were trained using the standard Leabra error-driven and associative (Hebbian) learning mechanisms to achieve target activations for every step in the sequence. In a more realistic model, the representations would not be predetermined but, rather, would develop as a function of the learning mechanisms; this shortcut was simply used as a convenient way of achieving a desired set of representations to test the basic sufficiency of our ideas about the gating mechanism.

For simplicity, every layer in the model has been organized into three different stripes, where a stripe corre-

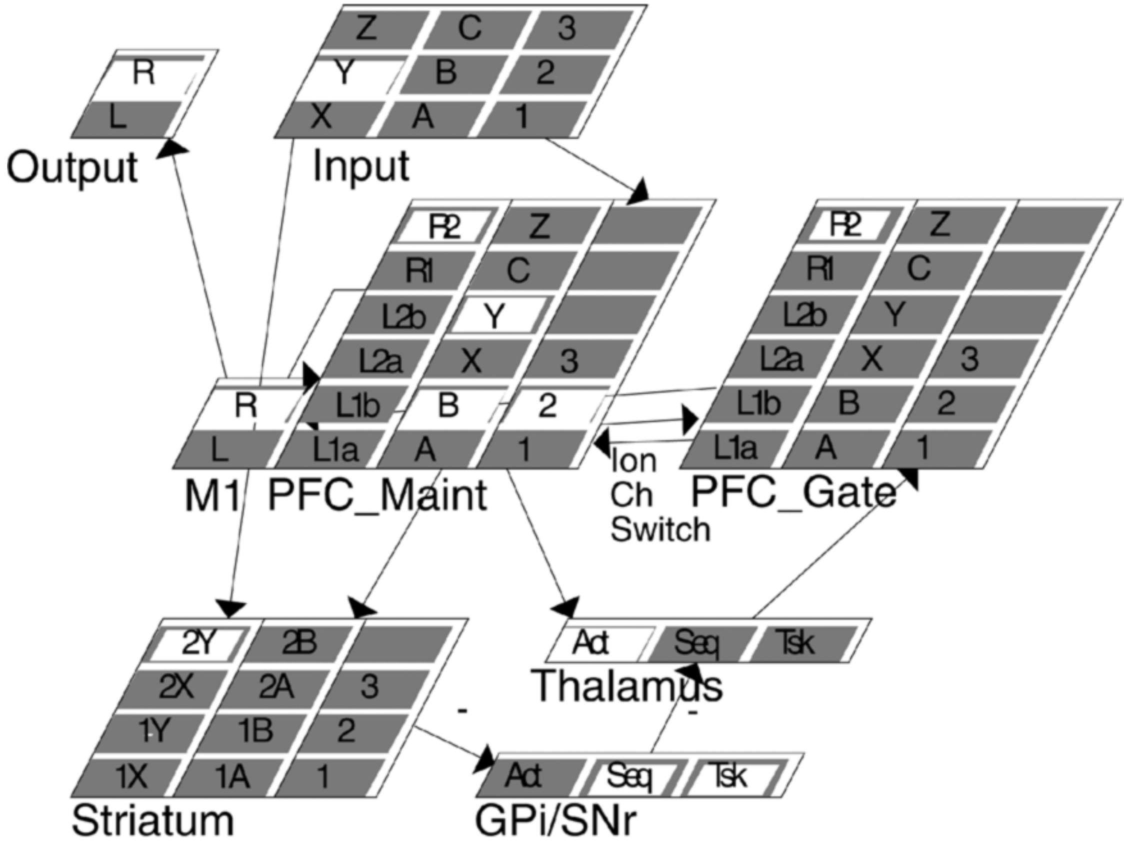

Figure 6. Working memory model with basal-ganglia-mediated selective gating mechanism. The network structure is analogous to Figure 4, but with the prefrontal cortex (PFC) has been subdivided into maintenance (PFC_Maint) and gating (PFC_Gate) layers. Three hierarchically organized *stripes* of the PFC and basal ganglia are represented as the three columns of units within each layer; each stripe is capable of being independently updated. The rightmost *task* (Tsk) stripe encodes task-level information (i.e., 1 or 2). The middle *sequence* (seq) encodes sequence-level information within a task (i.e., A or B). The leftmost *action* (act) stripe encodes action-level information (i.e., responding to the X or Y stimulus and actually producing the left or right output in PFC). Non-task-relevant inputs (e.g., 3, C, Z) are also presented, and the model ignores them—that is, they are not maintained.

sponds to an individually updatable region of the frontal cortex, as was discussed previously. The rightmost stripe in each layer represents the outer-loop task demand information (1 or 2). The middle stripe represents information maintained at the inner-loop, sequence level (A or B). The leftmost stripe represents stimuli that actually trigger an action response (X or Y). To clarify and simplify the motor aspects of the task, we have a response only at the end of an inner-loop sequence (i.e., after an X or Y), instead of responding L for all the preceding stimuli. All these other responses should be relatively automatic, whereas the response after the X or the Y requires taking into account all the information maintained in working memory, so it is really the task-critical motor response.

We describe the specific layers of the model (which match those shown in Figure 4, as was discussed previously) in the course of tracing a given trial of input. First, a stimulus is presented (activated) in the input layer. Every stimulus automatically activates its corresponding frontal representation, located in the PFC_Maint layer of the model. This layer represents cortical Layers 2–3 and 5–6 (without further distinguishing these layers, although it is possible there are divisions of labor between them) and is where stimulus information is represented and maintained. The other frontal layer is PFC_Gate, which represents the gating action of cortical Layer 4 (we will return to it in a moment).

If the input stimulus has been recognized as important for task performance, as a result of as-yet-unimplemented learning experiences (which are represented in the model through hand-set enhanced weight values), it will activate a corresponding unit in the striatum layer. This activation of the high-threshold striatal unit is the critical step in initiating the cascade of events that leads to maintaining stimuli in working memory, via a process of "releasing the brakes," or disinhibiting the thalamic loops through the frontal cortex. Note that these striatal units in the model encode *conjunctions* of maintained information in the frontal cortex (1 or 2, in this case) and incoming stimulus information (A, B, X, or Y). Although not computationally essential for this one task, these conjunctions reflect our theorizing that striatal neurons need to encode conjunctions in a high-threshold manner to avoid task-inappropriate stimulus activation. The frontal representations are also necessarily conjunctive in their detection of the combination of stimuli that trigger a response action; the stimulus maintenance representations could also be more conjunctive as well, even though it is not strictly necessary for this one task.

Once a striatal unit fires, it inhibits the globus pallidus unit in its corresponding stripe, which has, to this point, been tonically active and inhibiting the corresponding thalamus unit. Note the compression of the signal from the striatum to the globus pallidus, as was discussed above. The disinhibition of the thalamic unit opens up the recurrent loop that flows from the PFC_Maint units to the thalamus and back up to the PFC_Gate layer. Note that the disinhibited thalamic unit will only get activated if there is also descending activation from PFC_Maint units. Although this is always the case in our model, it would not be true if a basal ganglia stripe got activated (disinhibited) that did not correspond to an area of frontal activation; this property may be important for synchronizing frontal and basal ganglia representations during learning.

The effect of thalamic firing is to provide general activation to an entire stripe of units in the PFC_Gate layer. These frontal units cannot fire without this extra thalamic activation, but they also require excitation from units in the PFC_Maint layer, which are responsible for selecting the specific gate unit to activate. Although this is configured as a simple one-to-one mapping between maintenance and gating frontal units in the model, the real system could perform important kinds of learning here to fine-tune the gating mechanism. Finally, the activation of the gating unit controls the switchable excitatory ion channels in the frontal maintenance units. For those maintenance units within a stripe that receive both input from the current input stimulus and the gating activation, the excitatory ion channels are opened. Maintenance units that get only the gating activation, but not stimulus input, have their ion channels closed. This mechanism provides a means of updating working memory by resetting previously active units that are no longer receiving stimulus input, while providing sustained excitatory support for units that do have stimulus input.

An Example Sequence

Figure 7 shows an example sequence of 2–B–C–Y as processed by the model. The first stimulus presented is the task context—in this case, it is Task 2, the B–Y detection task. Because the striatum detects this stimulus as being task relevant (via the 2 striatal unit), it inhibits the task globus pallidus unit, which then disinhibits the corresponding thalamus unit. This disinhibition enables the thalamus to then become excited via descending projections from the frontal cortex. The thalamic activation then excites the PFC_Gate unit that also receives activation from the PFC_Maint layer, resulting in the activation of the excitatory ion channel for the 2 frontal unit in the PFC_Maint layer.

Next, the B input activates the 2B conjunctive striatal unit, which detects the combination of the 2 task maintained in the frontal cortex and the B stimulus input. This results in the firing of the sequence stripe and maintenance of the B stimulus encoding in the frontal cortex. Note that the 2 has been maintained as the B stimulus was being processed and encoded into active memory, owing to the fact that these items were represented in different stripes in the frontal cortex. This demonstrates the principle of selective gating, which is central to our model.

The next stimulus is a C distractor stimulus; this is not detected as important for the task by the striatum (i.e., all striatal units remain subthreshold) and is thus not gated into robust active maintenance (via the intrinsic ion channels). Note that despite this lack of gating, the C representation is still activated in the PFC_Maint frontal cortex layer, as long as the stimulus is present. However, when the next stimulus comes in (the Y in this case), the C activation decays quickly away.

Finally, the Y stimulus is important because it triggers an action. The 2Y striatal unit enables firing of the R2 unit in the PFC layers; this is a conjunctive unit that detects the conjunction of all the relevant working memory and input stimuli (2–B–Y, in this case) for triggering one kind of R output response (the other R conjunction would be a 1–A–X). This conjunctive unit then activates the basic R motor response, in a manner consistent with observed frontal recordings (e.g., Hoshi, Shima, & Tanji, 2000). Thus, the same basal-ganglia-mediated disinhibitory function supports both working memory updating and motor response initiation in this model.

Although it is not represented in this example, the model will maintain the 2 task signal over many inner-loop sequences (until a different task input is presented), because the inner-loop updating is selective and, therefore, does not interfere with maintenance of the outer-loop task information.

Summary

To summarize, the model illustrates how the frontal cortex can maintain information for *contextualizing* motor responses in a task-appropriate fashion, while the basal ganglia trigger the updating of these frontal representations and the initiation of motor responses.

DISCUSSION

We have presented a theoretical framework and an implemented neural network model for understanding how the frontal cortex and the basal ganglia interact in providing the mechanisms necessary for selective working memory updating and robust maintenance. We have addressed the following central questions in this paper.

1. What are the specific functional demands of working memory?

2. What is the overall division of labor between the frontal cortex and the basal ganglia in meeting these functional demands?

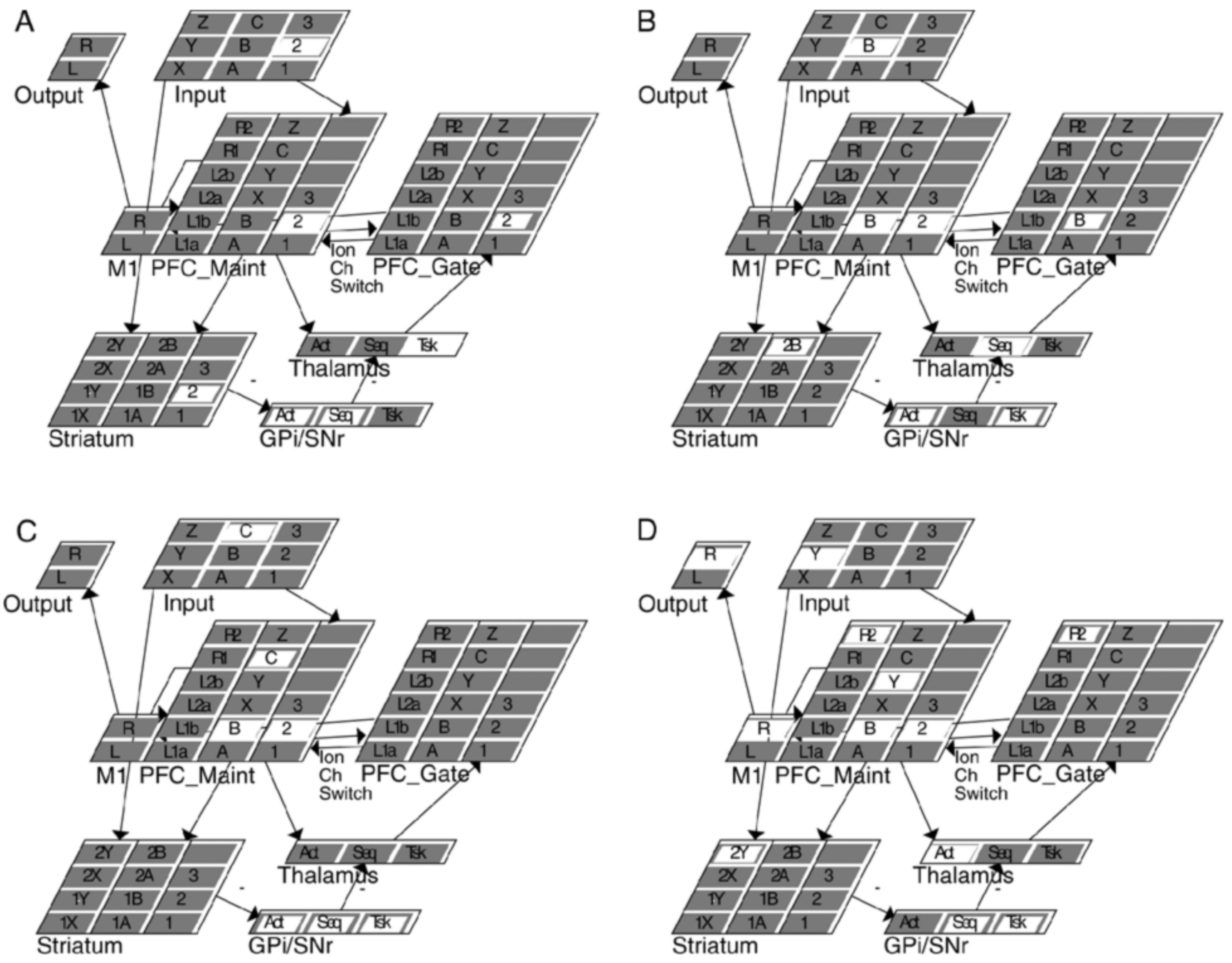

Figure 7. An example sequence in the model (2–B–C–Y). (A) Task Context 2 is presented. The striatum detects this stimulus as relevant and disinhibits the task stripe of the thalamus, allowing PFC_Gate to become active, causing the task number to be maintained in PFC_Maint. (B) The next stimulus is B, which the striatum detects in conjunction with Task Context 2 (from the PFC) via the 2B unit. The sequence stripe of the thalamus is then disinhibited, and B is gated into PFC_Maint, while Task Context 2 remains active owing to persistent ionic currents. This demonstrates *selective* gating. (C) A distractor stimulus C is presented, and because the striatum has not built up relevant associations to this stimulus, all units are subthreshold. The thalamus remains inhibited by the tonically active globus pallidus, and C is not maintained in the PFC. (D) Stimulus Y is presented, and the striatum detects the conjunction of it and the task context via the 2Y unit. The thalamus action level stripe is disinhibited, which activates conjunctive units in the frontal cortex (R2) that detect combinations of maintained and input stimuli (2–B–Y). These frontal units then activate the R response in the primary motor area (M1).

3. What kinds of specialized mechanisms are present in the frontal cortex to support its contributions to working memory?

4. What aspects of the complex basal ganglia circuitry are essential for providing its functionality?

Our answers to these questions are as follows.

1. Working memory requires robust maintenance (in the face of ongoing processing, other distractor stimuli, and other sources of interference), but also rapid, selective updating, where some working memory representations can be quickly updated while others are robustly maintained.

2. The frontal cortex provides maintenance mechanisms, whereas the basal ganglia provide selective gating mechanisms that can independently switch the maintenance mechanisms on or off in relatively small regions of the frontal cortex.

3. Frontal cortex neurons have intrinsic maintenance capabilities via persistent, excitatory ion channels that give maintained activation patterns the ability to persist without stimulus input. This allows frontal neurons to always encode stimulus inputs, while only maintaining selected stimuli, which is otherwise difficult using only recurrent excitatory attractor mechanisms. Recurrent connections play an additional maintenance role and are important for trial-and-error learning about what is important to maintain.

4. The disinhibitory nature of the basal ganglia effect on the frontal cortex is important for achieving a modu-

latory or gating-like effect. Striatal neurons must have a high effective threshold and selective, conjunctive representations (combining maintained frontal goal/task information with incoming stimuli) to fire only under specific conditions when updating is required. Although this conjunctivity requires large numbers of neurons, the striatal signal is collapsed down into a small number of globus pallidus neurons, consistent with the idea that the basal ganglia is important for determining *when* to do something, but not the details of what to do. The organization of this basal ganglia circuitry into a large number of parallel subcircuits, possibly aligned with the stripe structures of the frontal cortex, is essential for achieving a selective gating signal that allows some representations to be updated while others are maintained.

In the remaining sections, we discuss a range of issues, including the following: a comparison between our model and other theories and models in the literature; the unique predictions made by this model and, more generally, how the model relates to existing literature on the cognitive effects of basal ganglia damage; and the limitations of our model and future directions for our work.

OTHER THEORIES AND MODELS OF FRONTAL-CORTEX—BASAL-GANGLIA FUNCTION

We begin with an overview of general theories of the roles of the frontal cortex and basal ganglia system from a neuropsychological perspective and then review a range of more specific computational theories/models. We then contrast the present model with the earlier dopamine-based gating mechanisms.

General Theories

Our discussion is based on a comprehensive review of the literature on both the frontal cortex and the basal ganglia and on their relationship by Wise, Murray, and Gerfen (1996). They summarize the primary theories of frontal-cortex–basal-ganglia function according to four categories: attentional set shifting, working memory, response learning, and supervisory attention. We cover these theories in turn and then address some further issues.

Attentional set shifting. The attentional-set-shifting theory is supported in part by deficits observed from both frontal and basal ganglia damage—for example, patients perform normally on two individual tasks separately, but when required to switch dynamically between the two, they make significantly more errors than do normals (e.g., R. G. Brown & Marsden, 1990; Owen et al., 1993). This is exactly the kind of situation in which our model would predict deficits resulting from basal ganglia damage; indeed, the 1–2–AX task was specifically designed to have a task-switching outer loop because we think this specifically taps the basal ganglia contribution.

Furthermore, we and others have argued extensively that the basic mechanism of working memory function is integral to most of the cognitive functions attributed to the frontal cortex system (e.g., Cohen et al., 1996; Munakata, 1998; O'Reilly et al., 1999; O'Reilly & Munakata, 2000). For example, the robust maintenance capacity of the kinds of working memory mechanisms we have developed are necessary to maintain activations that focus attention in other parts of the brain on specific aspects of a task and to maintain goals and other task-relevant processing information. In short, one can view our model as providing a specific mechanistic implementation of the attentional-set-shifting idea (among other things).

This general account of how our model could address attentional-set-shifting data is bolstered by specific modeling work using our earlier dopamine-based gating mechanism to simulate the monkey frontal lesion data of Dias, Robbins, and Roberts (1997; O'Reilly, Noelle, Braver, & Cohen, 2001). The dopamine-based gating mechanism was capable of inducing task switching in frontal representations, so that damage to the frontal cortex resulted in slowed task switching. Moreover, we were able to account for the dissociation between dorsal and orbital frontal lesions observed by Dias et al. in terms of level of abstractness of frontal representations, instead of invoking entirely different kinds of processing for these areas. Thus, we demonstrated that an entirely working-memory-based model, augmented with a dynamic gating mechanism and some assumptions about the organization of frontal representations, could account for data that were originally interpreted in very different functional terms (i.e., attentional task shifting and overcoming previous associations of rewards).

Working memory. It is clear that our account is consistent with the working memory theory, but aside from a few papers showing the effects of caudate damage on working memory function (Butters & Rosvold, 1968; Divac, Rosvold, & Szwaracbart, 1967; Goldman & Rosvold, 1972), not much theorizing from a broad neuropsychological perspective has focused on the specific role of the basal ganglia in working memory. Thus, we hope that the present work will help to rekindle interest in this idea.

Response learning. This theory is closely associated with the ideas of Passingham (1993), who argued that the frontal cortex is more important for learning (specifically, learning about appropriate actions to take in specific circumstances) than for working memory. Certainly, there is ample evidence that the basal ganglia are important for reinforcement-based learning (e.g., Barto, 1995; Houk et al., 1995; Schultz et al., 1997; Schultz, Romo, et al., 1995), and we think that learning is essential for avoiding the (often implicit) invocation of a homunculus in theorizing about executive function and frontal control. However, we view learning within the context of the working memory framework. In this framework, frontal learning is about what information to maintain in an active state over time and how to update it in response to task demands. This learning should ensure that active representations have the appropriate impact on overall task performance, both by retaining useful information and by focusing attention on task-relevant information.

Supervisory attention. The supervisory attention theory of Norman and Shallice (Norman & Shallice, 1986; Shallice, 1988) is essentially that the *supervisory attention system* (SAS) controls action by modulating the operation of the *contention scheduling* (CS) system, which provides relatively automatic input/output response mappings. As is reviewed in Wise et al. (1996), other researchers (but not Shallice) have associated the SAS with the frontal cortex and the CS system with the basal ganglia. However, this mapping is inconsistent with the inability of the basal ganglia to directly produce motor output or other functions without frontal involvement; instead, as we and Wise et al. argue, the basal ganglia should be viewed as modulating the frontal cortex, which is the opposite of the SAS/CS framework.

Behavior-guiding rule learning. Wise et al. (1996) proposed their own overarching theory of the frontal-cortex–basal-ganglia system, which is closely related to other ideas (Fuster, 1989; Owen et al., 1993; Passingham, 1993). They propose that the frontal cortex is important for learning new *behavior-guiding rules* (which amount to sensory–motor mappings, in the simple case), whereas the basal ganglia modulate the application of existing rules as a function of current behavioral context and reinforcement factors. Thus, as in our model, they think of the basal ganglia as having a modulatory interaction with the frontal cortex. Furthermore, they emphasize that the frontal-cortex/basal-ganglia system should not be viewed as a simple motor control system but, rather, should be characterized as enabling flexible, dynamic behavior that coordinates sensory and motor processing. However, they do not provide any more specific, biologically based mechanisms for how this function could be carried out and how, in general, the frontal cortex and basal ganglia provide this extra flexibility. We consider our theory and model as an initial step toward developing a mechanistic framework that is generally consistent with these overall ideas.

Process versus content. The ideas above can all be characterized as describing different types of *processing* (switching, maintaining, modulating attention, learning). In contrast, theories of posterior cortex tend to focus on *content*, such as representing object identity ("what") in the ventral visual pathway versus spatial location ("where") or vision-for-action in the dorsal pathway. Can we also provide a content-based explanation of what is represented in the frontal-cortex/basal-ganglia system? In areas of the frontal cortex more directly associated with motor control, the content is obviously about muscles and sequences or plans of motor movements. In the monkey prefrontal cortex, attempts to incorporate the what/where distinctions from the posterior cortex (e.g., F. A. W. Wilson, Scalaidhe, & Goldman-Rakic, 1993) have been questioned (S. C. Rao et al., 1997), and the human neuroimaging data is also controversial (e.g., Nystrom et al., 2000). The original working memory theory of Baddeley (1986) posited a content-based distinction between visuospatial and verbal information, which has met with some support from neuroimaging studies showing a left/right (verbal/visuospatial) organization of frontal cortex (Smith & Jonides, 1997), although this is also not always consistently found (e.g., Nystrom et al., 2000).

As was mentioned previously, we have recently proposed that it might be more useful to think of the organization of frontal content in terms of level of abstractness (O'Reilly et al., 2001). Furthermore, we suggest that the frontal cortex represents a wide variety of content that is also encoded in the posterior cortex (but this content can be robustly maintained only in the frontal cortex), together with sensory–motor mappings, plans, and goals that may be uniquely represented in the frontal cortex (O'Reilly et al., 1999). Thus, more posterior and inferior areas have more concrete, specific representations, whereas more anterior and dorsal areas have more abstract representations. In the domain of motor control, it is known that more anterior areas contain progressively more abstract representations of plans and sequences—a similar progression can occur with more stimulus-based representations as well. This notion fits well with the ideas of Fuster (1989), who suggested that the frontal cortex sits at the top of a hierarchy of sensory–motor mapping pathways, where it is responsible for bridging the longest temporal gaps. This hierarchy can continue within the frontal cortex, with more anterior areas concerned with ever longer time scales and more abstract plans and concepts (see Christoff & Gabrieli, 2000, and Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999, for recent evidence consistent with this idea). The notion that the frontal cortex represents behavior-guiding rules (Wise et al., 1996) can be similarly fit within this overall paradigm by assuming that such rules are organized according to different levels of abstraction as well. Furthermore, the framework of Petrides (1994), which involves a distinction between simple maintenance versus more complex processing, can be recast as differences in levels of abstraction of the underlying representations (i.e., simple maintenance involves concrete representations of stimuli, whereas processing involves more abstract representations of goals, plans, and mappings).

Computational Theories/Models

Perhaps the dominant theme of extant computational models of the basal ganglia is that they support decision making and/or action selection (e.g., Amos, 2000; Beiser & Houk, 1998; Berns & Sejnowski, 1996; Houk & Wise, 1995; Jackson & Houghton, 1995; Kropotov & Etlinger, 1999; Wickens, 1993; Wickens et al., 1995; see Beiser, Hua, & Houk, 1997, and Wickens, 1997, for recent reviews). These *selection* models reflect a convergence between the overall idea that the basal ganglia are somehow important for linking stimuli with motor responses and the biological fact that striatal neurons are inhibitory and should, therefore, inhibit each other to produce selection effects. Thus, the basal ganglia could be important for selecting the best linkage between the current stimulus + context and a motor response, using inhibitory competition so that only the best match will "win." However, all of these theories suffer from the finding that striatal neurons do not appear to inhibit each other (Jaeger, Kita, & Wilson, 1994). One possible way of retaining this overall selection model is to have the inhibition work indirectly through dopamine modulation via the *indirect pathway* connections with the subthalamic nucleus, as was proposed by Berns and Sejnowski (1996). This model may be able to resolve some inconsistencies between the slice study that did not find evidence of lateral inhibition (Jaeger et al., 1994) and in vitro studies that do find such evidence (see Wickens, 1997, for a discussion of the relevant data); the slice preparation does not retain this larger scale indirect pathway circuitry, which may be providing the inhibition. This model is also generally consistent with the finding that dopamine appears to be important for establishing an independence of neural firing in the basal ganglia that is eliminated with damage to the dopamine system (Bergman et al., 1998).

At least some of the selection models also discuss the disinhibitory role of the basal ganglia and suggest that the end result is the initiation of motor actions or working memory updating (e.g., Beiser & Houk, 1998; Dominey, 1995). The selection idea has also been applied in the cognitive domain by simulating performance on the Wisconsin card sorting task (WCST; Amos, 2000). In this model, the striatal units act as match detectors between the target cards and the current stimulus and are modulated by frontal attentional signals. When an appropriate match is detected, a corresponding thalamic neuron is disinhibited, and this is taken as the network's response. Although this model does not capture the modulatory nature of the basal ganglia's impact on the frontal cortex and does not speak directly to the involvement of the basal ganglia in working memory, it nevertheless provides an interesting demonstration of normal and impaired cognitive performance on the WCST task, using the selection framework.

Our model is generally consistent with these selection models, insofar as we view the striatum as important for detecting specific conditions for initiating actions or updating working memory. As we emphasized earlier, this detection process must take into account contextual (e.g., prior actions, goals, task instructions) information maintained in the frontal cortex to determine whether a given stimulus is task relevant and, if so, which region of the frontal cortex should be updated. Thus, the basal ganglia under our model can be said to be performing the selection process of initiating an appropriate response to a given stimulus (or not).

Perhaps the closest model to our own is that of Beiser and Houk (1998), which is itself related to that of Dominey (1995) and is based on the theoretical ideas set forth by Houk and Wise (1995). This model has basal ganglia disinhibition resulting in the activation of recurrent cortico-thalamic working memory loops to maintain items in a stimulus sequence. Their maintenance mechanism involves a recurrent bistability in the cortico-thalamic loops, where a phasic disinhibition of the thalamus can switch the loop from the inactive to the active state, depending on a calcium channel rebound current. They also mention that the indirect path through the subthalamic nucleus could potentially deactivate these loops but do not implement this in the model. They apply this model to a simple sequence-encoding task involving three stimuli (A, B, C) presented in all possible orders. They show that for some parameter values, the network can spontaneously (without learning) encode these sequences, using unique activation patterns.

There are a number of important differences between our model and the Beiser and Houk (1998) model. First, as we discussed earlier, their use of recurrent loops for active maintenance incurs some difficulties that are avoided by the intracellular maintenance mechanisms employed in our model. For example, they explicitly separate the frontal neurons that encode stimulus inputs and those that maintain information, which means that their network would suffer from the catch-22 problem mentioned previously if they were to try to implement a learning mechanism for gating information into working memory. Furthermore, this separation constrains them to postulate direct thalamic activation resulting from striatal disinhibition (via the calcium rebound current), because the frontal neurons that project descending connections to the thalamus are not otherwise activated by stimuli. In contrast, our model has the thalamus being activated by descending frontal projections, as is consistent with available data showing that disinhibition alone is insufficient to activate the thalamus (Chevalier & Deniau, 1990).

Perhaps the most important difference is that their model does not actually implement a gating mechanism, because they do not deal with distractor stimuli, and it seems clear that their model would necessarily activate a working memory representation for each incoming stimulus. The hallmark of a true gating mechanism is that it selectively updates only for task-relevant stimuli, as defined by the current context. This is the reason that our model deals with a task domain that requires multiple, hierarchical levels of maintenance and gating, whereas their sequencing task requires only maintenance of the immediately prior stimulus, so that their model can succeed by always updating. Furthermore, their model has no provisions, or apparent need, for a learning mechanism, whereas this is a central, if presently incompletely implemented, aspect of our model.

To demonstrate that our model can also explain the role of the basal ganglia in sequencing tasks, we applied our architecture to a motor sequencing task that has been shown in monkeys to depend on the basal ganglia (Matsumoto, Hanakawa, Maki, Graybiel, & Kimura, 1999). In this task, two different sequences are trained—either 1, 2,

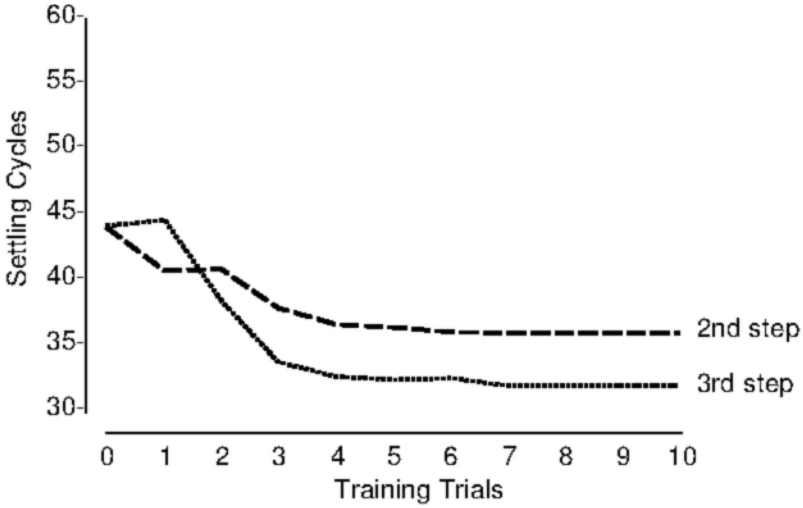

Figure 8. Settling time in the sequential network for the second and third steps in the sequence. The third step is faster owing to the ability of the network to predict it better.

3 or 1, 3, 2—where the numbers represent locations of lights in a display. Monkeys are trained to press buttons in the positions of these lights. The key property of these sequences is that after the second step in the sequence, the third step is completely predictable. Thus, it should be responded to faster, which is the case in intact monkeys, but not in monkeys with basal ganglia impairments. We showed that the network can learn to predict the third step in the sequence by encoding in the striatum a conjunction between the prior step and the onset of the third stimulus and, therefore, produce an output more rapidly. The same target representation kind of learning as that used in the 1–2–AX model was used to "train" this network. As a result of this learning's producing stronger representations, the network became better able to produce this prediction on the third step. This enabled the network to reproduce the basic finding from the monkey studies, a faster reaction time to the third step in the sequence (Figure 8). We anticipate being able to provide a more computationally satisfying sequencing model when we develop the learning aspect of our model in a more realistic fashion.

Finally, there are a number of other computational models that focus mainly on the PFC without a detailed consideration of the role of the basal ganglia (Dehaene & Changeux, 1989, 1991; Guigon, Dorizzi, Burnod, & Schultz, 1995; Moody et al., 1998; Seung, 1998; Tanaka & Okada, 1999; Zipser, 1991). Most of these models lack a true gating mechanism, even though some have a "gate" input that is additive and not modulatory, as required by a true gating mechanism (e.g., Moody et al., 1998). We argue that such models will suffer from the inability to dynamically shift between rapid updating and robust maintenance. For example, the Moody et al. model required 10 million training trials to acquire a simple delayed matching-to-sample task with distractors; we argue that this was because the network had to learn a delicate balance between updating and maintenance via additive network weights. This is consistent with computational comparisons of gated versus nongated memory models (Hochreiter & Schmidhuber, 1997). Other prefrontal-based models that were discussed earlier incorporate intrinsic bistability, as in our model (Durstewitz et al., 1999; Durstewitz et al., 2000a; Fellous et al., 1998; Wang, 1999), but these models lack explicit gating circuitry as implemented by the basal ganglia in our model.

To summarize, we see the primary contributions of the present work as linking the functional/computational level analysis of working memory function in terms of a selective gating mechanism with the underlying capacities of the basal-ganglia–frontal-cortex system. Although there are existing models that share many properties with our own, our emphasis on the gating function is novel. We have also provided a set of specific ideas, motivated again by functional/computational considerations, about active maintenance in terms of persistent ionic channels and how these could be modulated by the basal ganglia.

Relationship to the Dopamine-Based Gating Models

As was noted above, the present model was developed in the context of existing dopamine-based gating models of frontal cortex (Braver & Cohen, 2000; Cohen et al., 1996; O'Reilly et al., 1999; O'Reilly & Munakata, 2000). The primary difference between these models at the functional level is that the basal ganglia allow for *selective* updating, whereas dopamine is a relatively global neuromodulator that would result in updating large regions of the frontal cortex at the same time. In tasks that do not require this selective updating, however, we think that the two models would behave in a similar fashion overall. We will test this idea explicitly, after we have developed the learning mechanism for the basal ganglia model, by replicating earlier studies that have used the dopamine-based model.

Despite having a high level of overall functional similarity, these two models clearly make very different predictions regarding the role of dopamine in working memory. Perhaps the most important difference is that the dopamine-based gating mechanism is based on a coincidence between the need to gate information into working memory and differences in level of expected reward. Specifically, dopamine bursts are known to occur for unexpected rewards and, critically, for stimuli that have been previously predictive of future rewards (e.g., Montague et al., 1996; Schultz et al., 1993). Because it will, by definition, be rewarding to maintain stimuli that need to be maintained for successful task performance, it makes sense that dopamine bursts should occur for such stimuli (and computational models demonstrate that this is not a circular argument, even though it may sound like one; Braver & Cohen, 2000; O'Reilly & Munakata, 2000). However, this coincidence between reward prediction and the need to gate into working memory may not always hold up. In particular, it seems likely that after a task becomes well learned, rewards will no longer be unexpected, especially for intermediate steps in a chain of working memory updates (e.g., as required for mental arithmetic). The basal ganglia gating mechanism can avoid this problem because, in this model, dopamine is only thought to play a role in learning; after expertise is achieved, striatal neurons can be triggered directly from stimuli and context, without any facilitory boost from dopamine being required.

It remains possible that both dopamine and the basal ganglia work together to trigger gating. For example, broad, dopamine-based gating may be important during initial phases of learning a task, and then the basal ganglia play a dominant role for more well learned tasks. One piece of data consistent with such a scenario is the finding that, in anesthetized animals, dopamine can shift prefrontal neurons between two intrinsic bistable states (Lewis & O'Donnell, 2000). However, this finding has not been replicated in awake animals, so it is possible that normal tonic dopamine levels are sufficient to allow other activation signals (e.g., from Layer 4 activation driven by thalamic disinhibition) to switch bistable modes (as hypothesized in the present model). In short, further empirical work needs to be done to resolve these issues.

Unique Predictions and Behavioral Data

In addition to incorporating a wide range of known properties of the frontal cortex and basal ganglia system, our model makes a number of novel predictions at a range of different levels. At a basic biological level, the model incorporates a few features that remain somewhat speculative at this point, and therefore constitute clear predictions of the model that could be tested using a variety of electrophysiological methods:

1. Frontal neurons have some kind of intrinsic maintenance capacity—for example, excitatory ion channels that persist on the order of seconds. Note that subsequent predictions suggest that these currents will be activated only under very specific conditions, making them potentially somewhat difficult to find empirically.

2. Disinhibition of thalamic neurons should be a dominant factor in enabling the activation of corresponding Layer 4 frontal neurons.

3. Coactivation of Layer 4 neurons and other synaptic inputs into neurons in Layers 2–3 or 5–6 should lead to the activation of intrinsic maintenance currents. Activation of Layer 4 without other synaptic input should reset the intrinsic currents.

4. Frontal neurons within a stripe (e.g., within a short distance of each other, as in the isocoding columns of S. G. Rao et al., 1999) should all exhibit the same time course of updating and maintenance. For example, if one neuron shows evidence of being updated, others nearby should as well. Note that this does not mean that these neurons should necessarily encode identical information; different subsets of neurons within a stripe can be activated in different tasks or situations. However, the common gating signal should in general induce a greater level of commonality to neurons within a stripe than between.

At the behavioral level, the model has the potential to make detailed predictions regarding the different effects of lesions of each component along the circuit between the frontal cortex and the basal ganglia. At the most basic level, because these systems are mutually dependent, our model predicts that damage anywhere within the circuit will result in overall impairments (see L. L. Brown et al., 1997, R. G. Brown & Marsden, 1990, and Middleton & Strick, 2000b, for reviews of relevant data supporting this idea). For the more detailed predictions, we can only make qualitative predictions, because we have not directly modeled many behavioral tasks; we plan to use the learning-based version of our model (currently under development) to simulate a wide range of frontal tasks and to make more detailed predictions. The following are some suggestions of how damage to different parts of the model should differ in their behavioral consequences.

1. Selective damage to the basal ganglia (sparing the frontal cortex) should generally be more evident with more complex working memory tasks that require selective gating of information in the face of ongoing processing and/or other distracting information. Basic motor plans, sequences, and other kinds of frontal knowledge should remain intact. This suggestion is consistent with data reviewed in R. G. Brown and Marsden (1990), suggesting that Parkinson's patients show deficits most reliably when they have to maintain internal state information to perform tasks (i.e., working memory). For example, Parkinson's patients were selectively impaired on a Stroop task without external cues available, but not when these cues were available (R. G. Brown & Marsden, 1988). However, Parkinson's patients can also have reduced dopamine levels in the frontal cortex, so it is difficult to draw too many strong conclusions regarding selective basal gan-

glia effects from this population. Other evidence comes from neuroimaging studies that have found enhanced GPi activation in normals for a difficult planning task (Tower of London) and working memory tasks, but not in Parkinson's patients (Owen, Doyon, Dagher, Sadikot, & Evans, 1998). Another interesting case, with stroke-induced selective striatal damage and selective planning and working memory deficits, was reported by Robbins et al. (1995). They specifically interpreted this case as reflecting a deficit in the updating of strategies and working memory, which is consistent with our model.

2. Selective damage to the GPi will cause the frontal loops to be constantly disinhibited (which is presumably why pallidotomies are beneficial for enabling Parkinson's patients to move more freely). This should cause different kinds of behavioral errors, as compared with the effects of striatal damage, which would prevent the loops from becoming disinhibited. For example, constant disinhibition via GPi damage should result in excessive working memory updating according to our model, whereas striatal damage should result in an inability to selectively update at the appropriate time. Thus, GPi patients might appear scattered, impulsive, and motorically hyperactive, as in Huntington's syndrome, Tourette's syndrome, and people with attention-deficit hyperactivity disorder (all of which are known to involve the basal ganglia but have not been specifically linked with the GPi). In contrast, patients with striatal damage should exhibit both physical akinesia and "psychic akinesia" (R. G. Brown & Marsden, 1990)—the inability to initiate both actions and thoughts. Either updating too frequently or not enough will cause errors on many tasks (e.g., selective GPi damage has been described as impairing performance on a range of "frontal-like" tasks; Dujardin, Krystkowiak, Defebvre, Blond, & Destee, 2000; Trepanier, Saint-Cyr, Lozano, & Lang, 1998), but it should be possible to determine which of these problems is at work by analyzing the patterns of errors across trials.

3. Tasks that require multiple levels of working memory (e.g., the outer and inner loops of the 12–AX task) should activate different stripes in the frontal cortex, as compared with those that require only one level of working memory. Although it is entirely possible that these stripe-level differences would not be resolvable with present neuroimaging techniques, there is, in fact, some evidence consistent with this prediction. For example, an fMRI study has shown that activation is present in the anterior PFC specifically when "multitasking" is required (Christoff & Gabrieli, 2000; Koechlin et al., 1999). Other, more direct studies testing this prediction in the 12–AX are also currently underway, and preliminary data support our prediction (Jonathan D. Cohen, personal communication, February 1, 2001). One other possible experimental paradigm for exploring the model's predictions would be through the P300 component in event related potential studies, which has been suggested to reflect context updating (Donchin & Coles, 1988) and should be closely related to working memory updating (see also Kropotov & Etlinger, 1999).

Limitations of the Model and Future Directions

The primary limitation of our model as it stands now is in the lack of an implemented learning mechanism for shaping the basal ganglia gating mechanism so that it fires appropriately for task-relevant stimuli. In previous work, we and our colleagues have developed such learning models on the basis of the reinforcement learning paradigm (Braver & Cohen, 2000; Cohen et al., 1996; O'Reilly et al., 1999; O'Reilly et al., 2001). There is abundant motivation for thinking that the basal ganglia are intimately involved in this kind of learning, via their influence over the dopaminergic neurons of the substantia nigra pars compacta and the ventral tegmental area (e.g., Barto, 1995; Houk et al., 1995; Schultz et al., 1997; Schultz, Romo, et al., 1995).

The difficulty of extending this previous learning work to the present model comes from two factors. First, whereas in previous models we used a fairly abstract implementation of the dopaminergic system, we are attempting to make the new model faithful to the underlying biology of the basal ganglia system, about which much is known. Second, the selective nature of basal ganglia gating requires a mechanism capable of learning to allocate representations across the different separately controllable working memory stripes. In contrast, the earlier dopamine-based gating model had to contend only with one global gating signal. We are making progress addressing these issues in ongoing modeling work.

CONCLUSION

This research has demonstrated that computational models are useful for helping to understand how complex features of the underlying biology can give rise to aspects of cognitive function. Such models are particularly important when trying to understand how a number of different specialized brain areas (e.g., the frontal cortex and the basal ganglia) interact to perform one overall function (e.g., working memory). We have found in the present work a useful synergy between the functional demands of a selective gating mechanism in working memory and the detailed biological properties of the basal ganglia. This convergence across multiple levels of analysis is important for building confidence in the resulting theory.

REFERENCES

Alexander, G. E. (1987). Selective neuronal discharge in monkey putamen reflects intended direction of planned limb movements. *Experimental Brain Research*, **67**, 623-634.

Alexander, G. E., Crutcher, M., & DeLong, M. (1990). Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. In H. Uylings, C. Van Eden, J. De Bruin, M. Corner, & M. Feenstra (Eds.), *The prefrontal cortex: Its structure, function, and pathology* (pp. 119-146). Amsterdam: Elsevier.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, **9**, 357-381.

Amos, A. (2000). A computational model of information processing in the frontal cortex and basal ganglia. *Journal of Cognitive Neuroscience*, **12**, 505-519.

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

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215-232). Cambridge, MA: MIT Press.

Beiser, D. G., & Houk, J. C. (1998). Model of cortical-basal ganglionic processing: Encoding the serial order of sensory events. *Journal of Neurophysiology*, **79**, 3168-3188.

Beiser, D. G., Hua, S. E., & Houk, J. C. (1997). Network models of the basal ganglia. *Current Opinion in Neurobiology*, **7**, 185-190.

Bergman, H., Feingold, A., Nini, A., Raz, A., Slovin, H., Abeles, M., & Vaadia, E. (1998). Physiological aspects of information processing in the basal ganglia of normal and parkinsonian primates. *Trends in Neurosciences*, **21**, 32-38.

Berns, G. S., & Sejnowski, T. J. (1996). How the basal ganglia make decisions. In A. Damasio, H. Damasio, & Y. Christen (Eds.), *Neurobiology of decision-making* (pp. 101-113). Berlin: Springer-Verlag.

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.), *Attention and performance XVIII: Control of cognitive processes* (pp. 713-737). Cambridge, MA: MIT Press.

Brown, L. L., Schneider, J. S., & Lidsky, T. I. (1997). Sensory and cognitive functions of the basal ganglia. *Current Opinion in Neurobiology*, **7**, 157-163.

Brown, R. G., & Marsden, C. D. (1988). Internal versus external cues and the control of attention in Parkinson's disease. *Brain*, **111**, 323-345.

Brown, R. G., & Marsden, C. D. (1990). Cognitive function in Parkinson's disease: From description to theory. *Trends in Neurosciences*, **13**, 21-29.

Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: Emergent invariants and speed–accuracy properties during trajectory formation. *Psychological Review*, **95**, 49-90.

Butters, N., & Rosvold, H. E. (1968). The effect of caudate and septal nuclei lesions on resistance to extinction and delayed-alternation performance in monkeys. *Journal of Comparative Physiological Psychology*, **65**, 397-403.

Chevalier, G., & Deniau, J. M. (1990). Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*, **13**, 277-280.

Christoff, K., & Gabrieli, J. D. E. (2000). The frontopolar cortex and human cognition: Evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. *Psychobiology*, **28**, 168-186.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London: Series B*, **351**, 1515-1527.

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**, 332-361.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications* (pp. 267-296). Mahwah, NJ: Erlbaum.

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.

Constantinidis, C., & Steinmetz, M. A. (1996). Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. *Journal of Neurophysiology*, **76**, 1352-1355.

Dehaene, S., & Changeux, J. P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, **1**, 244-261.

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

Deniau, J. M., & Chevalier, G. (1985). Disinhibition as a basic process in the expression of striatal functions: II. The striato-nigral influence on thalamocortical cells of the ventromedial thalamic nucleus. *Brain Research*, **334**, 227-233.

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.

Dilmore, J. G., Gutkin, B. G., & Ermentrout, G. B. (1999). Effects of dopaminergic modulation of persistent sodium currents on the excitability of prefrontal cortical neurons: A computational study. *Neurocomputing*, **26**, 104-116.

Divac, I., Rosvold, H. E., & Szwaracbart, M. K. (1967). Behavioral effects of selective ablation of the caudate nucleus. *Journal of Comparative Physiological Psychology*, **63**, 184-190.

Dominey, P. F. (1995). Complex sensory–motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics*, **73**, 265-274.

Dominey, P. F., & Arbib, M. A. (1992). Cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, **2**, 153-175.

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral & Brain Sciences*, **11**, 357-427.

Douglas, R. J., & Martin, K. A. C. (1990). Neocortex. In G. M. Shepherd (Ed.), *The synaptic organization of the brain* (pp. 389-438). Oxford: Oxford University Press.

Dujardin, K., Krystkowiak, P., Defebvre, L., Blond, S., & Destee, A. (2000). A case of severe dysexecutive syndrome consecutive to chronic bilateral pallidal stimulation. *Neuropsychologia*, **38**, 1305-1315.

Durstewitz, D., Kelc, M., & Gunturkun, O. (1999). A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience*, **19**, 2807-2822.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000a). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology*, **83**, 1733-1750.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000b). Neurocomputational models of working memory. *Nature Neuroscience*, **3** (Suppl.), 1184-1191.

Erickson, S. L., & Lewis, D. A. (2000). Prefrontal cortical inputs to monkey mediodorsal thalamus. *Society for Neuroscience Abstracts* (p. 461). San Diego: Society for Neuroscience.

Fellous, J. M., Wang, X. J., & Lisman, J. E. (1998). A role for NMDA-receptor channels in working memory. *Nature Neuroscience*, **1**, 273-275.

Fox, C. A., & Rafols, J. A. (1976). The striatal efferents in the globus pallidus and in the substantia nigra. In M. D. Yahr (Ed.), *The basal ganglia* (pp. 37-55). New York: Raven.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, **61**, 331-349.

Fuster, J. M. (1989). *The prefrontal cortex: Anatomy, physiology and neuropsychology of the frontal lobe* (3rd ed.). New York: Lippincott-Raven.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, **173**, 652-654.

Gelfand, J., Gullapalli, V., Johnson, M., Raye, C., & Henderson, J. (1997). The dynamics of prefrontal cortico-thalamo-basal ganglionic loops and short term memory interference phenomena. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 253-258). Mahwah, NJ: Erlbaum.

Gobbel, J. R. (1995). A biophysically-based model of the neostriatum as a dynamically reconfigurable network. In M. Boden & L.-E. Niklasson (Eds.), *Proceedings of the Second Swedish Conference on Connectionism.* Hillsdale, NJ: Erlbaum.

Gobbel, J. R. (1997). *The role of the neostriatum in the execution of action sequences*. Unpublished doctoral dissertation, University of California, San Diego.

Goldman, P. S., & Rosvold, H. E. (1972). The effects of selective caudate lesions in infant and juvenile Rhesus monkeys. *Brain Research*, **43**, 53-66.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum & V. Mountcastle (Eds.), *Handbook of physiology*: *The nervous system* (Vol. 5, pp. 373-417). Bethesda, MD: American Physiological Society.

Goldman-Rakic, P. S., & Friedman, H. R. (1991). The circuitry of working memory revealed by anatomy and metabolic imaging. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 72-91). New York: Oxford University Press.

Gorelova, N. A., & Yang, C. R. (2000). Dopamine D1/D5 receptor activation modulates a persistent sodium current in rat prefrontal cortical neurons in vitro. *Journal of Neurophysiology*, **84**, 75-87.

Graybiel, A. M., & Kimura, M. (1995). Adaptive neural networks in the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 103-116). Cambridge, MA: MIT Press.

Guigon, E., Dorizzi, B., Burnod, Y., & Schultz, W. (1995). Neural correlates of learning in the prefrontal cortex of the monkey: A predictive model. *Cerebral Cortex*, **2**, 135-147.

Hikosaka, O. (1989). Role of basal ganglia in initiation of voluntary movements. In M. A. Arbib & S. Amari (Eds.), *Dynamic interactions in neural networks: Models and data* (pp. 153-167). Berlin: Springer-Verlag.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**, 1735-1780.

Hoshi, E., Shima, K., & Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*, **83**, 2355-2373.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233-248). Cambridge, MA: MIT Press.

Houk, J. C., & Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex*, **5**, 95-110.

Jackson, S., & Houghton, G. (1995). Sensorimotor selection and the basal ganglia: A neural network model. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 337-368). Cambridge, MA: MIT Press.

Jaeger, D., Kita, H., & Wilson, C. J. (1994). Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum. *Journal of Neurophysiology*, **72**, 2555-2558.

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature*, **399**, 148-151.

Kropotov, J. D., & Etlinger, S. C. (1999). Selection of actions in the basal ganglia-thalamocoritcal circuits: Review and model. *International Journal of Psychophysiology*, **31**, 197-217.

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, **34**, 337-347.

Lange, H., Thorner, G., & Hopf, A. (1976). Morphometric-statistical structure analysis of human striatum, pallidum, and nucleus subthalamicus: III. Nucleus subthalamicus. *Journal für Hirnforschung*, **17**, 31-41.

Levitt, J. B., Lewis, D. A., Yoshioka, T., & Lund, J. S. (1993). Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 & 46). *Journal of Comparative Neurology*, **338**, 360-376.

Lewis, B. L., & O'Donnell, P. (2000). Ventral tegmental area afferents to the prefrontal cortex maintain membrane potential "up" states in pyramidal neurons via D1 dopamine receptors. *Cerebral Cortex*, **10**, 1168-1175.

Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A. M., & Kimura, M. (1999). Role of nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of Neurophysiology*, **82**, 978-998.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

McFarland, N. R., & Haber, S. N. (2000). Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. *Journal of Neuroscience*, **20**, 3798-3813.

Middleton, F. A., & Strick, P. L. (2000a). Basal ganglia and cerebellar loops: Motor and cognitive circuits. *Brain Research Reviews*, **31**, 236-250.

Middleton, F. A., & Strick, P. L. (2000b). Basal ganglia output and cognition: Evidence from anatomical, behavioral, and clinical studies. *Brain & Cognition*, **42**, 183-200.

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**, 5154-5167.

Miyake, A., & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, **331**, 68-70.

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

Moody, S. L., Wise, S. P., di Pellegrino, G., & Zipser, D. (1998). A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. *Journal of Neuroscience*, **18**, 399-410.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the A-not-B task. *Developmental Science*, **1**, 161-184.

Neafsey, E. J., Hull, C. D., & Buchwald, N. A. (1978). Preparation for movement in the cat: I. Unit activity in the cerebral cortex. *Electroencephalography & Clinical Neurophysiology*, **44**, 714-723.

Norman, D., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. Davidson, G. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1-18). New York: Plenum.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *NeuroImage*, **11**, 424-446.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, **2**, 455-462.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 375-411). New York: Cambridge University Press.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, **4**, 661-682.

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. (2001). *Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control.* Manuscript submitted for publication.

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, **10**, 389-397.

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.

Owen, A. M., Doyon, J., Dagher, A., Sadikot, A., & Evans, A. C. (1998). Abnormal basal ganglia outflow in Parkinson's disease identified with PET: Implications for higher cortical functions. *Brain*, **121**, 949-965.

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.

Passingham, R. E. (1993). *The frontal lobes and voluntary action.* Oxford: Oxford University Press.

Petrides, M. (1994). Frontal lobes and working memory: Evidence from investigations of the effects of cortical excisions in nonhuman primates. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 9, pp. 59-82). Amsterdam: Elsevier.

Pucak, M. L., Levitt, J. B., Lund, J. S., & Lewis, D. A. (1996). Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *Journal of Comparative Neurology*, **376**, 614-630.

Rao, S. C., Rainer, G., & Miller, E. K. (1997, May). Integration of what and where in the primate prefrontal cortex. *Science*, **276**, 821-824.

Rao, S. G., Williams, G. V., & Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: Evidence for microcolumnar organization in PFC. *Journal of Neurophysiology*, **81**, 1903-1916.

Robbins, T. W., Shallice, T., Burgess, P. W., James, M., Rogers, R. D., Warburton, E., & Wise, R. S. J. (1995). Selective impairments in self-ordered working memory in a patient with a unilateral striatal lesion. *Neurocase*, **1**, 217-230.

Schneider, J. S. (1987). Basal ganglia-motor influences: Role of sensory gating. In J. S. Schneider & T. I. Lidsky (Eds.), *Basal ganglia and behavior: Sensory aspects of motor functioning* (pp. 103-121). Toronto: Hans Huber.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, **13**, 900-913.

Schultz, W., Apicella, P., Romo, R., & Scarnati, E. (1995). Context-dependent activity in primate striatum reflecting past and future behavioral events. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 11-28). Cambridge, MA: MIT Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997, March). A neural substrate of prediction and reward. *Science*, **275**, 1593-1599.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233-248). Cambridge, MA: MIT Press.

Seung, H. S. (1998). Continuous attractors and oculomotor control. *Neural Networks*, **11**, 1253-1258.

Shallice, T. (1988). *From neuropsychology to mental structure*. New York: Cambridge University Press.

Shima, K., & Tanji, J. (1998). Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. *Journal of Neurophysiology*, **80**, 3247-3260.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, **33**, 5-42.

Surmeier, D. J., & Kitai, S. T. (1999). D1 and D2 modulation of sodium and potassium currents in rat neostriatal neurons. *Progress in Brain Research*, **99**, 309-324.

Tanaka, S., & Okada, S. (1999). Functional prefrontal cortical circuitry for visuospatial working memory formation: A computational model. *Neurocomputing*, **26**, 891-899.

Taylor, J. G., & Taylor, N. R. (2000). Analysis of recurrent cortico-basal ganglia-thalamic loops for working memory. *Biological Cybernetics*, **82**, 415-432.

Trepanier, L. L., Saint-Cyr, J. A., Lozano, A. M., & Lang, A. E. (1998). Neuropsychological consequences of posteroventral pallidotomy for the treatment of Parkinson's disease. *Neurology*, **51**, 207-215.

Tsung, F.-S., & Cottrell, G. W. (1993). Learning simple arithmetic procedures. *Connection Science*, **5**, 37-58.

Wang, X.-J. (1999). Synaptic basis of cortical persistent activity: The importance of NMDA receptors to working memory. *Journal of Neuroscience*, **19**, 9587-9603.

Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research*, **89**, 233-247.

Wickens, J. [R.] (1993). *A theory of the striatum*. Oxford: Pergamon Press.

Wickens, J. [R.] (1997). Basal ganglia: Structure and computations. *Network: Computation in Neural Systems*, **8**, R77-R109.

Wickens, J. R., Kotter, R., & Alexander, M. E. (1995). Effects of local connectivity on striatal function: Simulation and analysis of a model. *Synapse*, **20**, 281-298.

Wilson, C. J. (1993). The generation of natural firing patterns in neostriatal neurons. In G. W. Arbuthnott & P. C. Emson (Eds.), *Chemical signalling in the basal ganglia* (Progress in Brain Research, Vol. 99, pp. 277-297). Amsterdam: Elsevier.

Wilson, F. A. W., Scalaidhe, S. P. O., & Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, **260**, 1955-1957.

Wise, S. P. (1985). The primate premotor cortex: Past, present, and preparatory. *Annual Review of Neuroscience*, **8**, 1-19.

Wise, S. P., Murray, E. A., & Gerfen, C. R. (1996). The frontal cortex-basal ganglia system in primates. *Critical Reviews in Neurobiology*, **10**, 317-356.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, **3**, 179-193.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, **13**, 3406-3420.

NOTES

[1] Other variations in target sequences for the two subtasks are possible and are being explored empirically.

[2] Although it is still possible that other frontal areas were really maintaining the signal during the intervening stimulus activations, this explanation becomes less appealing as this phenomenon is consistently observed across many different frontal areas.

APPENDIX
Implementational Details

The model is implemented using a subset of the Leabra framework (O'Reilly, 1998; O'Reilly & Munakata, 2000). The two relevant properties of this framework for the present model are (1) the use of a point neuron activation function and (2) the *k*-Winners-Take-All (*k*WTA) inhibition function that models the effects of inhibitory neurons. These two properties are described in detail below. In addition, the gating equations for modulating the intracellular maintenance ion currents in the PFC are described.

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 will be described below. Furthermore, use of 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:

$$\frac{dV_\mathrm{m}(t)}{dt} = \tau \sum_c g_c(t)\bar{g}_c\left[E_c - V_m(t)\right], \qquad (1)$$

APPENDIX (Continued)

with four channels (c) corresponding to the following: e, excitatory input; l, leak current; i, inhibitory input; and h for a hysteresis channel that reflects the action of a switchable persistent excitatory input; this h channel is used for the intracellular maintenance mechanism described below. 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 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:

$$\eta_j = g_e(t) \leq x_i w_{ij} \geq \frac{1}{n}\sum_i x_i w_{ij}. \tag{2}$$

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 χ:

$$y_j(t) = \frac{1}{\left(1 + \dfrac{1}{\gamma\left[V_m(t) - \Theta\right]_+}\right)}, \tag{3}$$

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 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, which reflects the intrinsic processing noise of biological neurons:

$$y_j^*(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/(2\sigma^2)} y_j(z - x)\,dz, \tag{4}$$

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 by using a numerical lookup table, since an analytical solution is not possible.

k-Winners-Take-All Inhibition

Leabra uses a kWTA function to achieve sparse distributed representations, with two different versions having different levels of flexibility around the k out of n active units constraint. Both versions compute a uniform level of inhibitory current for all units in the layer as follows:

$$g_i = g_{k+1}^{\Theta} + q\left(g_k^{\Theta} - g_{k+1}^{\Theta}\right), \tag{5}$$

where $0 < q < 1$ 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:

$$g_t^{\Theta} = \frac{g_e^* \overline{g}_e (E_e - \Theta) + g_l \overline{g}_l (E_l - \Theta)}{\Theta - E_i}, \tag{6}$$

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, g_k^{Θ} and g_{k+1}^{Θ} are set to the threshold inhibition value for the kth and k + 1th 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_t^{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 between .5 and .7, depending on the level of sparseness in the layer, with a standard default value of .6).

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.

APPENDIX (Continued)

Intracellular Ion Currents for PFC Maintenance

The gating function for switching on maintenance was implemented as follows: If any unit in the PFC Gating layer has activation that exceeds the *maintenance threshold*, the corresponding unit in the PFC Maintenance layer has its intracellular excitatory current (g_h) set to the value of the sending unit's (in PFC_Gate) activation, times the amount of excitatory input being received from the sensory input layer:

$$g_h = \begin{cases} x_i \eta_j & \text{if } x_i > \Theta_\mathrm{m} \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where x_i is the sending activation, η_j is the net input from the sensory input, and Θ_m is the maintenance threshold. If the g_h conductance is nonzero, it contributes a positive excitatory influence on the unit's membrane potential.

(Manuscript received November 22, 2000;
revision accepted for publication May 3, 2001.)