Prefrontal Cortex and Dynamic Categorization Tasks: Representational Organization and Neuromodulatory Control

Randall C. O'Reilly, David C. Noelle[1], Todd S. Braver[2] and Jonathan D. Cohen[3]

Department of Psychology, University of Colorado, Boulder, CO, [1]Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, [2]Department of Psychology, Washington University, St Louis, MO and [3]Department of Psychology, Princeton University, Princeton, NJ, USA

We present a computational model of the intradimensional/extradimensional (ID/ED) task (a variant of the Wisconsin card sorting task) that simulates the performance of intact and frontally lesioned monkeys on three different kinds of rule changes (Dias *et al.*, 1997, J Neurosci 17:9285–9297). Although Dias *et al.* interpret the lesion data as supporting a model in which prefrontal cortex is organized into different processing functions, our model suggests an alternative account based on representational content. A key aspect of the model is that prefrontal cortex representations are organized according to different levels of abstraction, with orbital areas encoding more specific featural information and dorsolateral areas encoding more abstract dimensional information. This representational scheme of the model is integrated with two additional key elements: (i) activation-based working memory representations controlled by a dynamic gating mechanism that simulates the hypothesized phasic actions of dopaminergic neuromodulation in prefrontal cortex, which acts to stabilize or destabilize frontal representations based on success in the task; and (ii) a weight-based associative learning system simulating posterior cortex and other subcortical areas, where the stimulus–response mappings are encoded. Frontal cortex contributes to the task via top-down activation-based biasing of task-appropriate features and dimensions in this posterior cortex system — this top-down biasing is specifically important for overcoming prepotent associations after a sorting rule reverses. The ability of the model to capture the double-dissociation observed by Dias *et al.* with orbital versus dorsolateral lesions supports the validity of these principles, many of which have also been useful in accounting for other frontal phenomena.

Introduction

The prefrontal cortex (PFC) is widely discussed as subserving a range of important functions in higher-level cognition, including executive control, planning, working memory, inhibition, decision making, and abstract thinking. These descriptive accounts help to make sense of behavioral data (e.g. from cases of prefrontal damage or neuroimaging results), but they provide little guidance in clarifying *how* the PFC achieves these various functions. One important way of understanding how the PFC works is to understand how its representations are organized. There have been a number of proposals regarding different representational schemes for frontal cortex, and considerable controversy still exists on this issue. This paper presents a mechanistically explicit computational model of PFC function that incorporates a novel proposal for the organization of frontal representations. This organizational scheme, together with a basic set of computational mechanisms (which apply throughout the frontal cortex), are used to account for an intriguing double-dissociation of impairments resulting from lesions of two different frontal areas on the ID/ED (intradimensional/extradimensional) dynamic categorization task (Dias *et al.*, 1997).

The ID/ED task (Roberts *et al.*, 1988; Owen *et al.*, 1993; Dias *et al.*, 1997) is a refinement of the widely studied Wisconsin card sorting task (WCST). Dynamic categorization tasks like these involve the periodic switching of response rules, with each rule in the ID/ED task defined as responding to a target stimulus in the context of several other stimuli. The ID/ED task involves two kinds of switches – intradimensional (ID) and extradimensional (ED). A dimension in this context represents a category of stimuli sharing the same general set of features (e.g. color, shape). For example, shapes constructed from homogeneous regions of color are considered one dimension, while shapes constructed from solid black lines are considered another dimension. An ID switch therefore involves changing the target stimulus to another within the same dimensional category (e.g. from one colored shape to another), while an ED switch involves changing the target stimulus to one from a different dimension (e.g. from a color shape to a solid black line figure). Dias *et al.* (Dias *et al.*, 1997) showed that orbital frontal lesions selectively impaired a particular type of ID switches, while dorsolateral frontal lesions selectively impaired specific ED switches.

These data clearly have the potential to inform the issue of how the PFC is organized, given that these two different frontal areas appear to be selectively involved in different aspects of the same task. One can categorize the existing proposals for understanding the organization of frontal cortex into at least two groups. One group, exemplified by Goldman-Rakic (Goldman-Rakic 1987), suggests that different areas of frontal cortex encode different kinds of representational *content* (e.g. spatial versus object representations), while performing the same kind of essential processing function (e.g. working memory). Another suggests that different frontal areas contribute qualitatively *different processing functions* [e.g. inhibition versus working memory (Fuster, 1989; Diamond, 1990) or simple maintenance versus complex processing (Petrides, 1994)]. Dias *et al.* (Dias *et al.*, 1997) interpreted their data as supporting a differential-processing model involving affective inhibition (in the orbital region) and attentional selection (in the dorsolateral region). In contrast, our model demonstrates that these findings are consistent with a content-based organization (with a common processing function of working memory across different areas).

Thus, our model shares the same basic approach as the working memory model of Goldman-Rakic (Goldman-Rakic, 1987), in that we view the single essential function of frontal cortex as that of maintaining information in an active state over time (i.e. *activation-based working memory*). Indeed, we have developed a set of biologically based computational mechanisms for understanding how the PFC is specialized for this active maintenance function (Cohen *et al.*, 1996; O'Reilly *et al.*, 1999; Braver and Cohen, 2000; Frank *et al.*, 2001). However, our model reflects a somewhat different view regarding the specific organization of representations within PFC. We take the view that an important principle of organization may be the level of

Cerebral Cortex Mar 2002;12:246–257; 1047–3211/02/$4.00

abstraction of representations, rather than the specific sensory modality or domain of the information represented. Thus, some areas may be responsible for representing specific featural information, while others are responsible for representing more abstract categories (e.g. featural dimensions) [see (Koechlin *et al.*, 1999; Christoff and Gabrieli, 2000) for related ideas].

The ID/ED Dynamic Categorization Task

The ID/ED task is a more complex version of the WCST, which has been widely used as a measure of frontal function [although performance on this task may also be affected by damage to other cortical areas (Stuss *et al.*, 2000)]. The ID/ED task has been explored in monkeys, neurologically intact humans, frontal patients, and Parkinson's patients (Owen *et al.*, 1993; Dias *et al.*, 1997; Roberts *et al.*, 1988). The task involves a two-alternative choice based on two stimuli that are presented on each trial (Fig. 1). Each choice stimulus is composed from two *features*, one from each of two different *dimensions* (e.g. filled-shape and line-shape). At any given point in the experiment, there are two different features from each dimension present (e.g. a bar and an X within the line-shapes dimension, and a triangle and square within the filled-shapes dimension). Therefore, there are four different possible choice stimulus configurations (Fig. 1). The correct choice is determined by one *target* feature from one dimension (e.g. the triangle feature from the filled-shape dimension), and the subject is rewarded for selecting the stimulus that contains the target. Thus, given only a single trial, reward is ambiguous between the two features of the stimulus on the rewarded side, but this ambiguity is resolved over multiple trials.

After the subject has achieved criterion level performance (90% correct), the target feature is changed (unbeknownst to the subject), and the four component stimuli either remain the same or change (Fig. 2). The target feature is changed either to a feature within the same dimension (*intradimensional*, or ID), or to a feature in the other dimension (*extradimensional*, or ED). When the stimuli remain the same, the rule change results in a *reversal*, because the selection of a previously rewarded feature must be reversed to become a non-selection, and the non-selection of a previously unrewarded feature must be reversed to become a selection. When the stimuli change, the rule change results in a *shift*, because the selection of a previously rewarded feature must be shifted onto a novel feature. Thus, there are four different types of changes to the categorization rule: intradimensional reversal (IDR), intradimensional shift (IDS), extradimensional reversal (EDR) and extradimensional shift (EDS).

It is important to emphasize that although the EDS case is a shift at the level of the features (new features are used), it is really a reversal at the level of the dimensions involved. This is because prior to the EDS, subjects learn that one dimension is relevant and the other is not. Then, the EDS requires these dimensions to be reversed – the previously irrelevant dimension must be reversed to become relevant, and a previously relevant dimension must be reversed to become irrelevant. The EDR condition, which was not run in the Dias *et al.* study, is therefore a reversal at both the featural (like IDR) and dimensional levels (like EDS).

In the Dias *et al.* study (Dias *et al.*, 1997), marmosets with lesions in either dorsolateral or orbital (ventromedial) frontal cortex were tested on three of these changes: IDR, IDS and EDS. They found selective impairment of IDRs with orbital lesions, selective impairment of EDS with dorsolateral lesions, and no significant impairments on IDSs with either frontal lesion (Fig. 3).

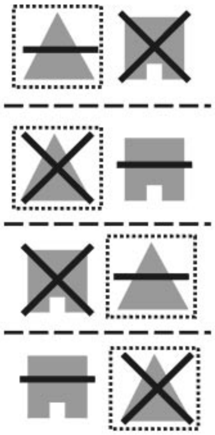

Figure 1. Example stimuli from the feature and dimension switching categorization task studied by Dias *et al*. (Dias *et al*., 1997). Each row represents one trial. Subjects were rewarded for choosing the target (indicated by the surrounding box, which was not presented to the subjects). The stimuli were created by combining one of two possible line-shapes and one of two possible filled-shapes (all four possible test combinations are shown for these line-shape and filled-shape dimensions). The target was determined by one feature from one dimension, in this case, the triangle feature from the filled-shape dimension.

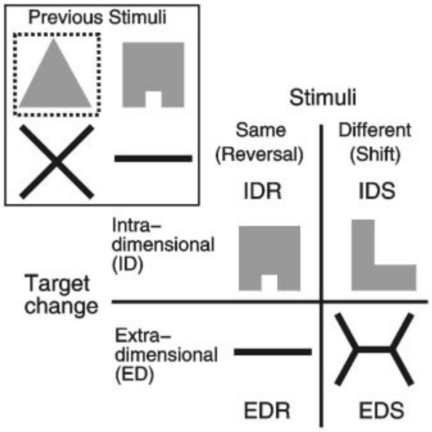

Figure 2. Types of changes in categorization rule following criterion performance on the original rule (Fig. 1), organized by same versus different stimuli across the rule change, and intradimensional or extradimensional target change. The target stimulus is again indicated by the surrounding box.

Dias *et al.* (Dias *et al.*, 1997) interpreted their results as follows:

> It suggests that distinct regions of the prefrontal cortex carry out independent but complementary forms of cognitive processing of complex visual stimuli in changing environmental circumstances. Thus, regions within the orbital prefrontal cortex in marmosets enable the rapid reversal of affective associations for specific visual stimuli, whereas the higher-order shifting of attention between supraordinate features of visual stimuli, such as their perceptual dimensions, is mediated by regions within the lateral prefrontal cortex. (p. 9296).

As we noted in the Introduction, this way of characterizing the different frontal areas is consistent with functionally based

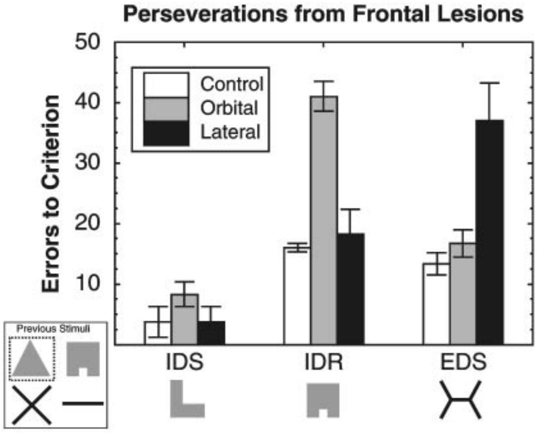

Figure 3. Perseverations from different types of frontal lesions in the WCST-like task studied by Dias *et al*. (Dias *et al*., 1997), reflected in number of errors made before reaching training criterion. Orbital lesions (ventromedial) selectively impair IDRs, whereas lateral lesions selectively impair EDSs (which are actually reversals to a previously ignored dimension). IDSs are never impaired. Data from Dias *et al*. (Dias *et al*., 1997).

theories (Fuster, 1989; Diamond, 1990; Petrides, 1994), although it also does assume a representational distinction between 'specific visual stimuli' and 'supraordinate features'. Dias *et al*. (Dias *et al*., 1997) argued that their results were inconsistent with a working memory ('on-line processing') account, because the apparent memory demands across the different tasks (IDR, IDS, EDS) are all roughly equivalent, whereas deficits were observed only when inhibition or higher-order dimensional shifting was required. We argue that this analysis is based on a simplified notion of working memory that suggests it is only important for the raw maintenance of information. In contrast, we think that the maintenance of information in frontal cortex, when combined with representational distinctions along the lines suggested by Dias *et al*. (Dias *et al*., 1997), has important implications for processing elsewhere in the system, in ways that can explain the Dias *et al*. pattern of results.

A Theory of Perseverations in the ID/ED Task

First, to avoid confusion arising from the many different uses of the term *working memory*, we restrict our usage to refer to the process of maintaining information through persistent neural firing (i.e. *activation-based* working memory), which has long been observed in frontal cortex (Fuster and Alexander, 1971; Kubota and Niki, 1971). The primary advantage of this form of working memory for the present purposes is that it can be rapidly updated by simply changing the activation state of a set of neurons. In contrast, more long-term memories encoded in connection weights between neurons require structural changes to update, which can be much slower (but more enduring). Therefore, activation-based memories support more *flexible* processing in the sense that a variety of different strategies, rules, or goals can be quickly activated and de-activated.

We argue that this kind of flexibility is essential for rapid switching in dynamic categorization tasks. Specifically, we think the PFC maintains a representation of the currently relevant dimension or feature in activation-based working memory, and that this activation provides top-down support or *biasing* (Cohen and O'Reilly, 1996) of the corresponding perceptual processing and action selection pathways, facilitating sorting along this dimension or feature. Categorization behavior may be altered either as a result of weight changes in these modulated pathways, or through changes in the activation-based biasing provided by the working memory system. While the active representation can be relatively rapidly switched when the sorting rules change, a weight-based solution must slowly unlearn the previous weights and learn the new ones. Thus, the perseveration observed in the ID/ED and WCST tasks can be accounted for by the loss of the more flexible, prefrontally mediated activation-based working memory, causing behavior to depend solely on the less flexible weight-based learning supported by posterior cortex and other brain areas [see (Munakata, 1998; Munakata *et al*., 2001) for similar ideas].

Importantly, the effects of frontal damage are only evident when there are prepotent responses that must be overcome – this is when the frontal top-down support of the new target (dimension or feature) is critical for overcoming the strength of the prior target. Such prepotent responses are present only after the first rule is learned in WCST, and in ID/ED only in the conditions of IDR and EDS (which involves a reversal of the relevant dimension, as noted previously), but not IDS (which involves all new stimuli and no reversal of the relevant dimension). The fact that activation-based top-down support interacts with the strength of existing associations means that an assessment of the role of working memory in this task based on the idea that it only maintains information [as provided by Dias *et al*. (Dias *et al*., 1997)] is incomplete. Maintained representations influence processing elsewhere in the brain, and this influence is felt in some conditions more than others, depending on the strength of learned associations in those other areas. Although it may be intuitively appealing to describe such conditions as requiring *inhibition of prepotent responses*, we can provide a more parsimonious overall account of frontal function by thinking instead in terms of sustained activations *supporting weaker responses* (Cohen and O'Reilly, 1996; Munakata, 1998; O'Reilly and Munakata, 2000; Miller and Cohen, 2001) [cf. the biased-competition model of Desimone and Duncan (Desimone and Duncan, 1995)].

The flexibility conferred by PFC is specific to the representational content of PFC areas. We suggest that these areas are organized along a gradient of abstraction, such that: (i) orbital frontal cortex in the marmoset is particularly important for supporting feature-level representations, so that lesions here impair reversals at the feature level (IDR); and (ii) dorsolateral areas in the marmoset support dimension-level representations, so that lesions here impair reversals at the dimension level (EDS). As noted, these representational distinctions were also assumed by Dias *et al*. (Dias *et al*., 1997). We discuss possible extensions of this account to other primate species in the Discussion.

The Dynamic Gating Mechanism

The flexibility of activation-based working memories depends critically on the presence of a *dynamic gating* mechanism, which controls the updating and maintenance of working memory representations. When the gate is open, working memory can be updated, and when it is closed, any currently active working memories are protected from interference (from noise, ongoing processing, irrelevant stimuli, etc.). This gate is needed because one setting of connection strengths into the working memory system cannot satisfy both the need for rapid updating and robust maintenance (Cohen *et al*., 1996; O'Reilly *et al*., 1999; Braver and Cohen, 2000; O'Reilly and Munakata, 2000). Biologically, we have shown how this gating mechanism can be implemented either through the phasic dopamine neuromodulation of the frontal cortex by the ventral tegmental area (VTA) (Durstewitz *et al*., 1999; Braver and Cohen, 2000;

O'Reilly and Munakata, 2000) (which is used in the present model), or through the interactions between the basal ganglia and frontal cortex (Frank *et al.*, 2001). In either case, the updating properties of the gating mechanism are shaped by a reinforcement-based learning mechanism, which plays a critical role in the present model by triggering the updating of working memory representations when the categorization rule changes.

Basic Properties of the Model

Figure 4 shows the structure of the model, which is implemented using the Leabra framework (O'Reilly, 1998, 2001; O'Reilly and Munakata, 2000) (see Appendix). Leabra integrates a number of widely used neural network mechanisms into one coherent framework, and has been used to simulate over 40 different psychological phenomena in O'Reilly and Munakata (O'Reilly and Munakata, 2000). Thus, all the basic mechanisms in the model have strong independent motivation; we note below where specific features of the algorithm play an important role.

The input layer represents the stimuli in a simple format, with separate units for the two different dimensions in each of the two different locations (left and right). There are four units within each dimension, and features are encoded using simple distributed representations having two out of the four units active. The posterior cortex layer, which represents the sensory cortex encoding of the stimuli, is organized in the same way as the input, but is limited so that it can only have two units active at the same time, so that once a given target feature has been learned, the posterior cortex representations of the other features are naturally suppressed. This reflects attentional limitations in processing that have been shown to exist in the visual system (Desimone and Duncan, 1995) – an entire complex multi-stimulus input pattern cannot be processed in parallel. This constraint is implemented via the k-Winners-Take-All mechanism in Leabra, described in the Appendix. The output response is produced via connections directly from the posterior cortex layer, meaning that all of the actual response outputs are generated via weight-based associations between posterior cortex and output units, and not by any direct outputs from frontal cortex. Instead, as argued earlier, the frontal cortex contributes via top-down biasing of posterior cortex. Therefore, we restricted synaptic learning to those connections involving the posterior cortex but not the frontal cortex (and the VTA inputs as described below), so that we could ensure that frontal cortex was contributing through an activation-based mechanism and not through synaptic modification. In reality, both types of contributions would be possible.

The PFC is organized into two areas: orbital areas that represent featural information, and lateral areas that represent more abstract dimension-level information. These areas are reciprocally interconnected with the posterior cortex units, and their activity thus biases the posterior cortex units. The featural nature of the orbital representations is accomplished by having the individual PFC units connect in a one-to-one fashion with the featural units within the two dimensions. We only included one set of four such featural units per dimension so that this area does not encode the location of the features, only their identity. The dimensional nature of the lateral representations comes from the fact that there are only two units in this layer, one for each dimension, with each unit fully connected with the feature units in the posterior cortex layer from the corresponding dimension. Throughout, the network connectivity generally obeys the principle of bidirectional cortical connections (Felleman and Van Essen, 1991; O'Reilly and Munakata, 2000). The exceptions are the VTA which projects to the PFC (see below), the dimensional PFC which projects to the featural PFC but not vice versa, in accord with the ideas and data reviewed by Gobbel (Gobbel, 1997) [see also Frank *et al.* (Frank *et al.*, 2001)], and the input layer which is fixed and therefore does not receive backprojections.

During initial learning, the network has no difficulty activ-

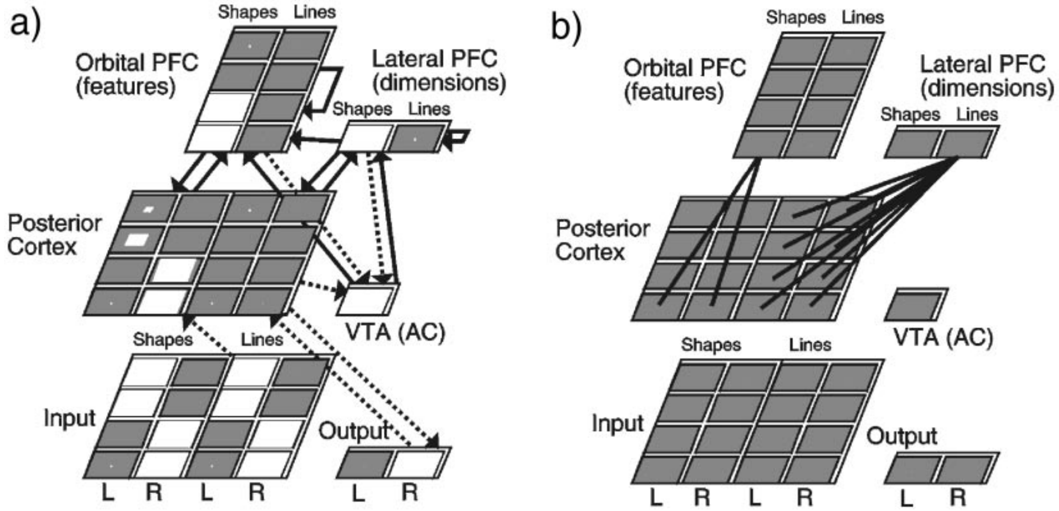

Figure 4. (*a*) Dynamic categorization network with typical activation state (higher activation = larger white square): the input contains two stimuli (left = L and right = R), and the network must choose one. The stimuli differ along two dimensions (shapes and lines); features within a dimension are represented by two active units within a column. The Input and Posterior Cortex layers encode these features separately. Response (Output) is generated by learned associations from the posterior cortex (weights subject to learning are shown with dotted lines). The posterior cortex layer cannot represent all of the input features simultaneously; learning activates relevant features while suppressing irrelevant ones. Thus, learning is impaired when the irrelevant features become relevant. The PFC layers (representing feature-level and dimension-level information) help by more rapidly focusing posterior cortex layer processing on previously irrelevant features. The VTA acts like an adaptive critic (AC), stabilizing and destabilizing PFC units in response to errors. Connectivity is bidirectional except VTA only controls PFC, and the more abstract dimensional PFC unidirectionally projects to featural PFC (input is fixed and thus receives no back-projection). (*b*) Each dimensional (lateral) PFC unit projects reciprocally to the entire set of hidden units representing that dimension; the featural (orbital) PFC units project reciprocally to individual sets of features. Thus, dimensional PFC is more effective at dimensional switching even though featural units are also dimension-specific.

ating the target item representation because all posterior cortex units are roughly equally likely to get activated, and the correct item will get reinforced through learning (both the Hebbian and error-driven learning mechanisms in Leabra will cause the target item representation to become stronger). However, if the target is then switched to one that was previously irrelevant (i.e. a reversal), then the irrelevant item will not tend to be activated in the posterior cortex layer, making it difficult to learn the new association. The top-down PFC biasing can overcome this problem by supporting the activation of the new target item, giving it a competitive edge in the limited activation competition.

The VTA layer represents the ventral tegmental area, which provides a dynamic gating mechanism via dopamine neuromodulation to the PFC. It has been shown that the VTA fires dopamine bursts for stimuli that are predictive of reward (Schultz *et al.*, 1993), in a way that is generally consistent with the temporal differences reinforcement learning mechanism (Sutton, 1988; Montague *et al.*, 1996). If rewards are expected but not delivered (i.e. due to a behavioral error), the dopamine neurons exhibit reduced firing, corresponding to a *negative error signal*. We reasoned that task-relevant information that should be maintained is a reliable predictor of reward, and should thus elicit dopamine firing, resulting in the updating of working memory (O'Reilly *et al.*, 1999; Braver and Cohen, 2000), and that the negative error signal should reset working memory representations. The net effect is to produce a form of *trial-and-error search* by activating and deactivating PFC representations.

The VTA unit directly modulates the strength of weights in the PFC according to *changes* in its activity. When VTA transitions from not expecting reward to getting a reward (0 to 1), the weights from the posterior cortex units to the PFC transiently increase, thereby encoding the current pattern of posterior cortex activity. If an error is made after correct performance, the negative change in expected reward (1 to 0) causes the PFC gating to decrease significantly in strength, including the gain on the recurrent self-maintenance weights. This effectively clears the PFC activations. If there is no change in the VTA activation, the PFC will maintain its current values. There is also noise so random activation of PFC can occur, especially when there is nothing already in the PFC. Detailed equations are presented in the Appendix.

The model was trained and tested following the procedures of Dias *et al.* (Dias *et al.*, 1997). The network received two blocks of training followed by the rule change. The first block had only features from one dimension (one of which was the target), and the second block added the two features from the other dimension (but the target remained the same). Each block was trained until a criterion of two epochs (passes through all training items) without error. In the third block, the categorization rule changed as an IDS, IDR, EDS or EDR, and we measured the number of epochs needed to learn the new rule. Three types of networks were run: intact, feature-level (orbital PFC) lesion, and dimension-level (lateral PFC) lesion. Lesions were implemented by effectively removing the corresponding PFC layer. Ten different networks with random initial weights were run.

Results

The results from the model (Fig. 5) capture the essential features of the monkey data. Featural (orbital PFC) lesions cause selective deficits on IDRs, while dimensional (lateral PFC) lesions cause selective deficits on extradimensional shift/reversals (EDS). Neither type of lesion affects the IDSs.

The general explanation for these results, which is substantiated in detail next, is that the dopamine-based gating control mechanism produces rapid switching of PFC representations when expected rewards are not received (i.e. when errors are made just after the rule is changed). Although both areas of PFC have the same functional characteristics, they differ in the nature of their preexisting representations, which leads to the selectivity of the deficit. The more abstract dimensional representations in lateral PFC selectively affect extradimensional switching (producing deficits on EDS if lateral PFC is lesioned), while the more concrete featural representations in orbital PFC selectively affect intradimensional switching (producing deficits on IDR if orbital PFC is lesioned). Because an IDR stays within the same dimension, lateral PFC lesions have no effect on this condition, and conversely, orbital PFC lesions have no effect on EDS because this involves a dimensional, not featural, reversal. In both IDR and EDS cases, it is only when there is a competing

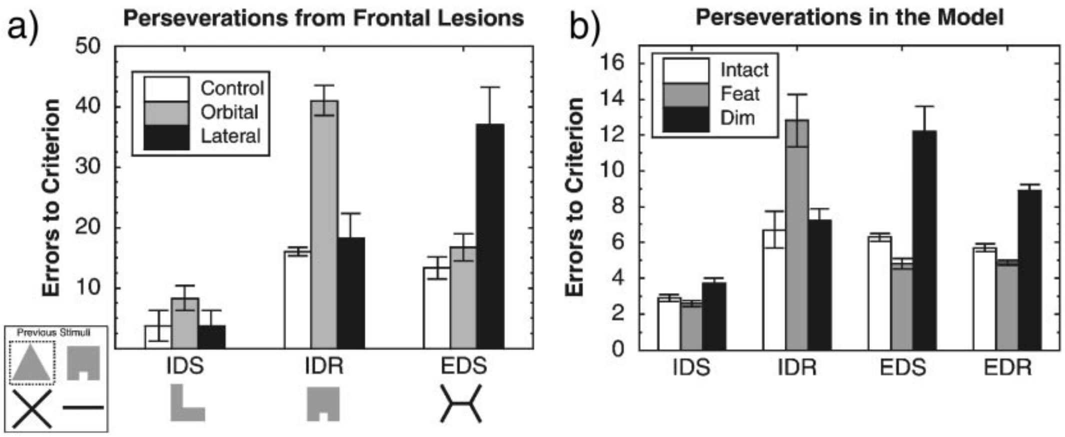

Figure 5. Perseverations from different types of simulated frontal lesions in the model (*b*), showing number of errors made before criterion performance was reached, as compared with data from monkeys shown previously in Figure 3 (*a*). *Feat* is lesions of the feature-level PFC representations, which correspond to the orbital lesions (ventromedial) in the Dias *et al*. (Dias *et al*., 1997) monkeys — intradimensional reversals (IDR) are selectively impaired. *Dim* is lesions of the dimension-level PFC representations, which correspond to the lateral lesions in monkeys — extradimensional shift/reversals (EDS) are selectively impaired. IDSs are never impaired. A novel prediction of the model is that dimensional (lateral) lesions should selectively impair EDRs. Error bars represent standard error of the mean across 10 different simulation runs.

prepotent response (i.e. in conditions involving some form of reversal) that switching deficits are observable – otherwise, only relatively small weight changes are needed to produce effective learning in the posterior system.

Note that even though different orbital PFC units encode different dimensions (Fig. 4*b*), and could therefore help to switch to another dimension, there are two properties that work against this. First, as shown in Figure 4*b*, the dimensional units project reciprocally to the entire set of features within a dimension, and are thus much more effective at switching across dimensions than the featural representations that only project to a small subset of a dimension. In short, *any* different activation in the dimensional layer will produce a switch to a new dimension, whereas many different activations in the featural layer will not shift across dimensions. Second, the overlap among features within a given dimension produces a bias towards activating other featural PFC units within the same dimension.

To illustrate this explanation in specific cases, Figure 6*a* shows the state of the network after initial target acquisition. The target is the triangle filled shape, and this is what is activated in posterior cortex and the orbital PFC, with the shape dimension active in lateral PFC. Then, in the IDR condition (Fig. 6*b*), the input stimuli are the same, but the network's initial responses (based on the old triangle target) are wrong, which rapidly destabilizes the PFC representations. When the PFC activates the other shape representation (squarish shape), this provides top-down support for the corresponding representation in posterior cortex, which then facilitates shifting the responding to be based on this target. In the EDS condition (Fig. 6*c*), the input stimuli change to all new items, and responding is initially random but still focused on the old, wrong dimension (shape instead of lines). As errors occur, the PFC representations are destabilized, and when the new dimension (lines) is activated in lateral PFC, this provides top-down support to that dimension in posterior cortex, facilitating learning of the correct output mapping. The detailed time course of this lateral PFC-mediated switching is shown for one EDS run in Figure 7, where it is clear that the dimensional unit switching leads to switching in the other layers.

One small but interesting difference between the model and monkey data is that the featural (orbital) lesions appear to *improve* extradimensional shift performance (EDS) in the model. This effect can be attributed to the fact that top-down activation from the featural level of the PFC can hinder dimensional reversals, *impairing* performance slightly unless this portion of PFC is lesioned. Although this effect makes sense in the model, it is likely that other collateral effects of damage, or a less perfect division of dimensional and featural representations, could obscure such a small effect in the monkeys. Also, the model learns overall faster than the monkeys (i.e. it has lower overall errors to criterion) – this can be attributed to the small size of the network and the fact that it is completely focused on this task. Although we could potentially have slowed down the model's learning, we chose instead to use standard learning rate parameters.

The results of the EDR condition constitute an interesting prediction, showing that only dimensional (lateral) PFC lesions cause impairments, despite the fact that it is like the IDR condition in having the same stimuli present. The notion that orbital PFC is responsible for inhibiting affective associations associated with stimuli, as proposed by Dias *et al*. (Dias *et al*., 1997), would appear to predict that orbital PFC lesions would impair performance in this EDR condition. Instead, we suggest that because the dimensional PFC units do not encode featural level details anyway, they will be unaffected by having the same stimuli present. Thus, their role in switching in EDR should be similar to that in EDS, as is suggested by the model results.

Discussion

This simulation demonstrates that an activation-based working memory model, combined with frontal representations organized according by different levels of abstractness, can account for the double-dissociation of orbital and dorsolateral frontal lesions observed by Dias *et al*. (Dias *et al*., 1997). Specifically, we found that by simulating the role of dopamine in regulating the frontal cortex in terms of an adaptive-critic mechanism, a rapid trial-and-error searching process emerged. This searching process deactivated the PFC when errors were made, and

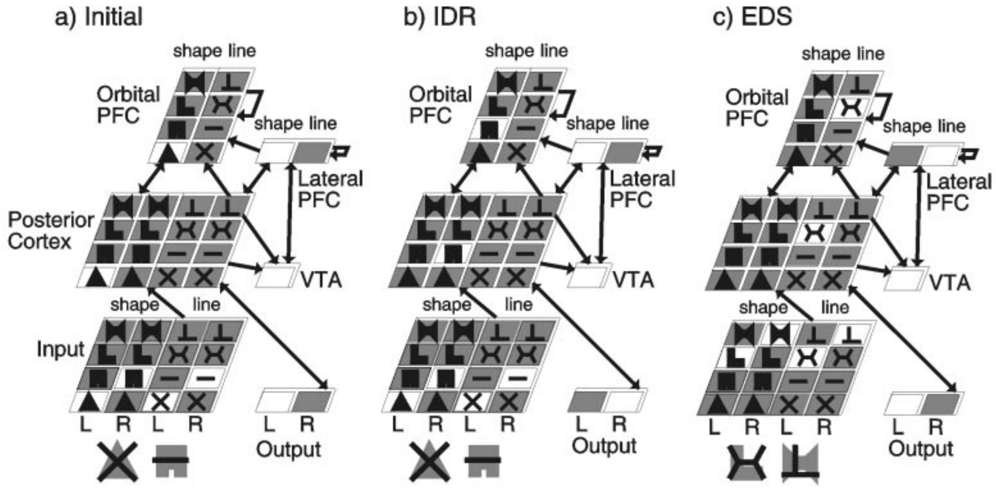

Figure 6. Schematic of ID/ED network performance using localist representations instead of the distributed ones present in the actual network (active units shown in white). (*a*) The state of the network after learning the initial problem (target = triangle-like filled shape). (*b*) The network after undergoing an IDR, where the target is now the box-like shape. The orbital (feature) PFC units switch rapidly under activation-based control, and then provide top-down support for the corresponding representation in the posterior cortex. (*c*) The network after undergoing an EDS, where the new target is the new X-like line shape (with a horizontal bar in the middle). Again, the PFC units rapidly switch, this time triggered by the lateral (dimensional) area, facilitating attention for the units in the new line dimension.

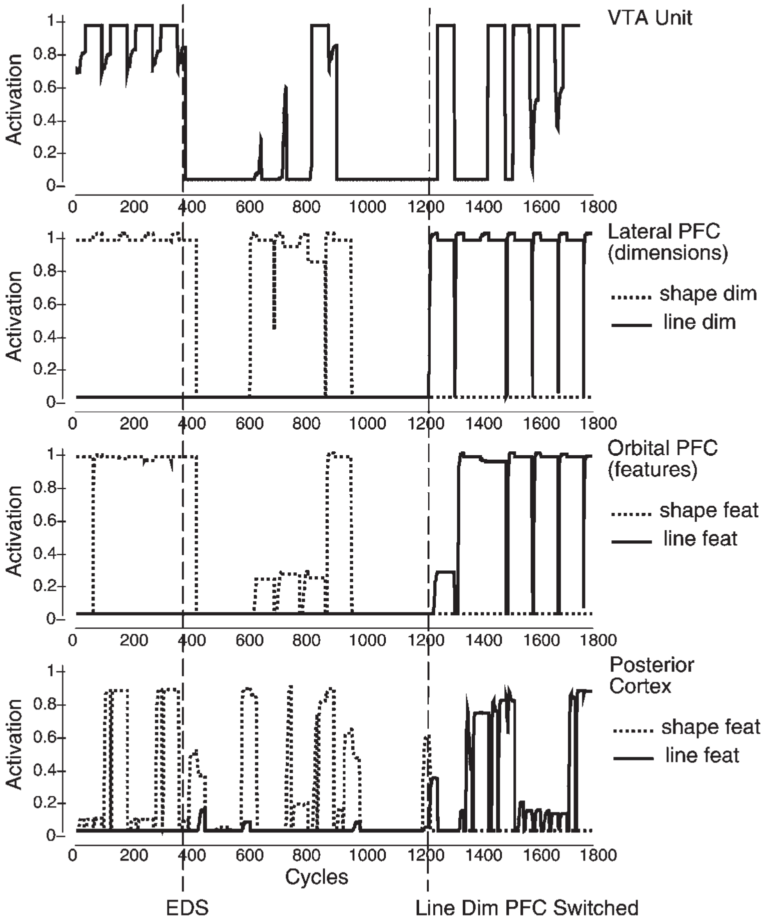

Figure 7. Time course of activations for key units in the VTA (dopamine gating), lateral (dimensional) PFC, orbital (featural) PFC, and posterior cortex layers of the intact model during an EDS trial (dimensional reversal). Prior to the EDS, shape-coding units in each layer are appropriately active (dotted lines), and the VTA unit is consistently predicting rewards (i.e. it is consistently active). Shortly after the onset of the EDS, the network makes errors, which results in deactivation of the VTA and consequently a negative dopamine change, causing both PFC layers to become deactivated via the dopamine-based gating mechanism. Although the shape units are subsequently reactivated for a bit, the resulting errors shut them down again, and then the lateral PFC unit for the line dimension becomes activated, which produces good performance by causing line representations to become activated in orbital PFC and posterior cortex. The resulting good performance causes the VTA to become activated (producing a positive dopamine burst), and this stabilizes the PFC representations. Note that while the PFC units are constantly active, the posterior units plotted are only relevant for a subset of trials, as is evident in the graph. Trials are 60 cycles long (one cycle = one activation update; see Appendix).

activated it either through noise or when performance was successful. The model provided a qualitative match to the effects of orbital and lateral PFC damage in marmosets on the dynamic categorization task by encoding more detailed feature-level information in orbital areas, while encoding more abstract dimension-level information in lateral areas. These different levels of representation, when combined with the trial-and-error control mechanism, provided a quick way of reconfiguring the categorization rule used by the network via top-down biasing of different posterior representations. This biasing effect conferred selective benefits on performance when the rules were reversed, in which case the cortical system had a difficult time overcoming the prior (dominant) pattern of responding without the help of top-down activity.

At a more general level, this model provides an important step toward characterizing the kinds of control mechanisms that enable activation-based working memory to confer greater flexibility as compared to weight-based adaptations. Although the activation-based working memory framework instantiated in our model is consistent with other active or working memory ideas (e.g. Goldman-Rakic, 1987), it extends these notions by accounting for empirical data that implicates the frontal cortex in overcoming perseverations through greater dynamic flexibility.

There are other important advantages of activation-based working memories (O'Reilly and Munakata, 2000). Specifically, working memory is useful 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 (Miyake and Shah, 1999). Furthermore, as exploited in the present model, the active nature of working memory provides a natural mechanism for *cognitive control* (or *task-based attention*), where top-down activation can influence processing elsewhere to achieve task-relevant objectives (Cohen *et al.*, 1990; Cohen and O'Reilly, 1996; O'Reilly *et al.*, 1999). Thus, working memory and cognitive control can be seen as two different sides of the same coin of actively maintained information.

However, these advantages of activation-based working memories also have concomitant disadvantages (O'Reilly and Munakata, 2000). 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 – for example the dynamic gating mechanisms needed for rapid updating and robust maintenance in frontally mediated activation-based working memory. This type of tradeoff-based reasoning has also been exploited in the context of weight-based long-term episodic memory (O'Reilly and Rudy, 2001).

Our model also provides support for the general principle that different areas of frontal cortex might be organized according to different levels of abstractness. In the marmoset, we argued that this organization exists along the ventral (more specific) to lateral (more abstract) axis. However, these organizational axes may differ in other species. For example, it is possible that the lateral–ventral axis in monkeys is more closely aligned with the anterior–posterior axis in humans. Therefore, we want to emphasize instead the general principle that different frontal areas may be organized according to different levels of abstraction, without making specific claims as to the correspondence of the marmoset brain areas with those in other primates. With this idea in mind, we can explore how well abstraction might align with other interpretations of frontal organization in more familiar primate species.

In the rhesus macaque monkey, a number of neural recording studies suggest that more ventral areas encode more specific object or pattern information (Mishkin and Manning, 1978; Wilson *et al.*, 1993). Other researchers have hypothesized, based on a variety of data, that the dorsolateral areas in humans are involved in more complex, abstract processing, whereas the ventral areas are used for simpler memory processes that require maintaining specific information (Petrides, 1994). In contrast to this generally dorsal–ventral organization, others have suggested an anterior–posterior organization where anterior prefrontal areas have the most temporally-extended representations (Koechlin *et al.*, 1999). We can interpret this temporal extension as a form of abstraction (Frank *et al.*, 2001; O'Reilly *et al.*, 1999), where more anterior/dorsal frontal areas could encode information that is relevant over longer delays (i.e. more abstract goals or plans that encompass many subgoals or subplans), while more posterior/ventral areas encode information relevant over briefer intervals. A major project for future work is to explore how much of this other data can be accounted for using explicit computational models incorporating the principle of differing levels of abstraction.

There are other aspects of the Dias *et al.* data (Dias *et al.*, 1997) that the model cannot address in its present form. Specifically, the monkeys were also tested on additional reversals and shifts, with the result that the second reversal did not produce the same patterns of frontal deficits as the first. We suspect that this result emerges as a result of the posterior system establishing a more equal balance among the representations involved, and thus do not consider it a challenge to the basic principles captured in our present model. Simulating this data would require a more complex network capable of representing at least 16 different stimuli, whereas our current model only handles four.

More generally, there are a number of scaling issues raised by the model – we used a small number of units to simulate a complex phenomenon that likely involves millions of neurons in the monkey brain. This is one of a number of challenges that are often leveled at neural network models of this sort, and a full discussion is beyond the scope of this paper [see O'Reilly and Munakata (O'Reilly and Munakata, 2000) for one such discussion]. One of the most important ways of addressing such concerns is to ask, 'will the fundamental principles behind the model's behavior change with scaling?'. In this case, we think not – the most basic principles of the differences between weight- versus activation-based memories should not depend critically on scaling parameters, and we can understand the model's behavior in terms of these principles. Thus, we must be careful to think of the model as just one possible concrete implementation of more general principles.

There are various aspects of the model that could be improved with future research. For example, we have hand-coded the frontal representations in this model (which therefore do not have learned connections as shown in Fig. 4), but it should be possible for these representations to develop naturally through learning in response to a combination of initial architectural constraints and task demands. Demonstrating this kind of learning is important for avoiding hidden 'homunculi' in the model, and is therefore a topic of active research in our group.

Predictions

The model and the broader theory in which it is framed make a number of testable predictions. For example, we simulated the EDR condition in the model [which was not run by Dias *et al.* (Dias *et al.*, 1997)], and found that only lateral (dimensional) PFC lesions impaired performance on this condition. Orbital (featural) PFC lesions had no apparent effect, despite the fact that this condition involves all the same stimuli, and therefore appears to require 'inhibiting affective associations' with the previous target item [which is how Dias *et al.* (Dias *et al.*, 1997) interpreted the role of the orbital PFC]. We interpreted this finding as showing that for the network, the dimensional switching represents the dominant difficulty for the EDR case, and that because the dimensional PFC units are relatively abstract anyway, their ability to switch is relatively unaffected by featural-level changes. This finding thus represents a testable prediction from the model, and one that appears to distinguish it from the predictions that Dias *et al.* (Dias *et al.*, 1997) would make.

At a more general level, the overall framework behind the model makes a number of predictions. For example, we predict that electrophysiological recordings would reveal differences in the extent to which neurons in different areas of the PFC exhibit abstract, categorical representations of stimuli. This could be tested by using an experiment similar to that performed by Freedman *et al.*, who have shown that prefrontal neurons in the macaque encode abstract categories of stimuli such as cats versus dogs (Freedman *et al.*, 2001).

Similarly, neuroimaging studies in monkeys or humans could be used to test our ideas. For example, as mentioned above, we have suggested that neuroimaging studies showing activation of more anterior PFC areas in humans (Koechlin *et al.*, 1999; Christoff and Gabrieli, 2000) can be interpreted as reflecting an organization according to different levels of abstraction. Assuming such an organization, our framework would predict that one might be able to find evidence of a anterior–posterior organization for IDR versus EDS activation in the ID/ED task in humans. Indeed, one such study found suggestive evidence consistent with this prediction (Rogers *et al.*, 2000). They found that when EDS and IDR were directly compared, there was more dorsal/anterior PFC activation during EDS, but no more ventral PFC activity during IDR. The lack of a ventral difference for IDR versus EDS suggests that these regions were active for both IDR and EDS, which is consistent with their non-shifting control comparison as well.

Other Models

Although we are not aware of any other computational models of the ID/ED task specifically, there are several published models that share some features with the present model. First, the hypothesized role of frontal cortex in our model is very similar to that proposed by Cohen *et al.* in their model of the Stroop task (Cohen *et al.*, 1990). They suggested that frontal cortex is specifically important for overcoming the prepotent process of word reading to support the weaker process of color naming. We have subsequently generalized this Stroop model in terms of top-down biasing as a mechanism by which the frontal cortex contributes to *cognitive control* or *controlled processing* (Cohen and O'Reilly, 1996; Miller and Cohen, 2001). A key feature of dynamic categorization tasks such as ID/ED is that frontal representations need to be rapidly updated when the rules change to provide useful top-down activation-based support, whereas in the Stroop model the frontal representations were externally specified (e.g. through task instructions). Thus, the present model extends the Stroop model by demonstrating how dynamic gating mechanisms for activation-based working memory can provide rapidly switched, task-appropriate top-down biasing.

The Dehaene and Changeux model (Dehaene and Changeux, 1991) of the WCST has some similarities to our own, in that it is based on an error-driven search mechanism. However, a critical difference is that their model relies on weight-based learning and unlearning that is modulated by the error signals, and is not fundamentally a top-down biasing based model like the one presented here. Specifically, when an error occurs, negative weights are incremented to prevent the return to a previously unsuccessful sorting rule. Indeed, despite the presence of sustained activation memory units in their model, their primary 'memory' manipulation involved changing the decay parameter of these negative weights. To simulate a frontal lesion, they changed the strength of weights into the error unit that is responsible for setting these negative weights, instead of damaging their sustained-firing memory units. Thus, the Dehaene and Changeux model (Dehaene and Changeux, 1991) makes very different mechanistic assumptions, and is less closely tied to specific brain areas, than the model presented here.

Levine and Prueitt (Levine and Prueitt, 1989) also presented a WCST model that is generally similar to the Dehaene and Changeux (Dehaene and Changeux, 1991) model, but lacks the decay function on the error-driven negative weights and some other features. Moving outside the realm of neural network models, Kimberg and Farah presented a production-system framework that accounted for a range of frontal deficits, including perseveration on the WCST (Kimberg and Farah, 1993). The essence of the model is that frontal damage reduces the influence of specific information on production firing, such that the productions end up falling back on perseverative and noisy firing biases that operate in the absence of other specific information. Thus, they build in perseveration as the behavior that the model resorts to after a frontal lesion. In contrast, we see perseveration as a result of learning in the weight-based processing of the posterior cortex. Nonetheless, this paper makes a number of more general points that resonate well with the framework presented here. For example, Kimberg and Farah emphasize the idea that frontal cortex can be understood as performing a single function, that, when damaged, produces a range of different behavioral manifestations (Kimberg and Farah, 1993). Furthermore, this common frontal function has something generally to do with working memory, which is consistent with our framework.

Notes

We thank Yuko Munakata for helpful comments. This work was supported in part by NIH grant MH47566 (RO and JC), NSF grant IBN-9873 492 (RO), and ONR grant N00014-00-1-0246 (RO). A preliminary version of this model appeared in O'Reilly and Munakata (O'Reilly and Munakata, 2000).

Address correspondence to Randall C. O'Reilly, Department of Psychology, University of Colorado Boulder, 345 UCB, Boulder, CO 80309, USA. Email: oreilly@psych.colorado.edu.

References

Braver TS, Cohen JD (2000) On the control of control: the role of dopamine in regulating prefrontal function and working memory. In: Control of cognitive processes: attention and performance XVIII (Monsell S, Driver J, eds), pp. 713–737. Cambridge, MA: MIT Press.

Christoff K, Gabrieli JDE (2000) The frontopolar cortex and human cognition: evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. Psychobiology 28:168–186.

Cohen JD, O'Reilly RC (1996) A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In: Prospective memory: theory and applications (Brandimonte M, Einstein GO, McDaniel, MA, eds), pp. 267–296. Mahwah, NJ: Erlbaum.

Cohen JD, Braver TS, O'Reilly RC (1996) A computational approach to prefrontal cortex, cognitive control, and schizophrenia: recent developments and current challenges. Phil Trans R Soc Lond B Sci 351:1515–1527.

Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing model of the Stroop effect. Psychol Rev 97:332–361.

Dehaene S, Changeux JP (1991) The Wisconsin Card Sorting Test: theoretical analysis and modeling in a neuronal network. Cereb Cortex 1:62–79.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193.

Diamond A (1990) The development and neural bases of memory functions as indexed by the A-not-B task: evidence for dependence on dorsolateral prefrontal cortex. In: The development and neural bases of higher cognitive functions (Diamond A, ed.), pp. 267–317. New York: New York Academy of Science Press.

Dias R, Robbins TW, Roberts AC (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. J Neurosci 17:9285–9297.

Durstewitz D, Kelc M, Gunturkun O (1999) A neurocomputational theory of the dopaminergic modulation of working memory functions. J Neurosci 19:2807.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Frank MJ, Loughry B, O'Reilly RC (2001) Interactions between the frontal cortex and basal ganglia in working memory: a computational model. Cognit Affect Behav Neurosci 1:137–160.

Freedman D, Riesenhuber M, Poggio T, Miller E (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316.

Fuster JM (1989) The prefrontal cortex: anatomy, physiology and neuropsychology of the frontal lobe. New York: Raven Press.

Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173:652–654.

Gobbel JR (1997) The role of the neostriatum in the execution of action sequences. PhD thesis, University of California, San Diego, San Diego, CA.

Goldman-Rakic PS (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. Handbook of physiology — the nervous system, vol. 5, pp. 373–417.

Kimberg DY, Farah MJ (1993) A unified account of cognitive impairments following frontal lobe damage: the role of working memory in complex, organized behavior. J Exp Psychol Gen 122:411–428.

Koechlin E, Basso G, Grafman J (1999) The role of the anterior prefrontal cortex in human cognition. Nature 399:148.

Kubota K, Niki H (1971) Prefrontal cortical unit activity and delayed alternation performance in monkeys. J Neurophysiol 34:337–347.

Levine DS, Prueitt PS (1989) Modeling some effects of frontal lobe damage — novelty and perseveration. Neural Netw 2:103–116.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Mishkin M, Manning FJ (1978) Nonspatial memory after selective prefrontal lesions in monkeys. Brain Res 143:313–323.

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

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neuroscience 16:1936–1947.

Munakata Y (1998) Infant perseveration and implications for object permanence theories: a PDP model of the *AB* task. Dev Sci 1:161–184.

Munakata Y, Morton JB, Stedron JM (2001) The role of prefrontal cortex in perseveration: developmental and computational explorations. In: Connectionist models of development (Quinlan P, ed.). Hove: Psychology Press, in press.

O'Reilly RC (1996) Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. Neural Comput 8:895–938.

O'Reilly RC (1998) Six principles for biologically-based computational models of cortical cognition. Trends Cogn Sci 2:455–462.

O'Reilly RC (2001) Generalization in interactive networks: the benefits of inhibitory competition and Hebbian learning. Neural Comput 13:1199–1242.

O'Reilly RC, Munakata Y (2000) Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. Cambridge, MA: MIT Press.

O'Reilly RC, Rudy JW (2001) Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol Rev 108:311–345.

O'Reilly RC, Braver TS, Cohen JD (1999) A biologically based computational model of working memory. In: Models of working memory: mechanisms of active maintenance and executive control (Miyake A, Shah P, eds), pp. 375–411. New York: Cambridge University Press.

Owen AM, Roberts AC, Hodges JR, Summers BA, Polkey CE, Robbins TW (1993) Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. Brain 116:1159–1175.

Petrides M (1994) Frontal lobes and working memory: evidence from investigations of the effects of cortical excisions in nonhuman primates. In: Handbook of neuropsychology, vol. 9 (Boller F, Grafman J, eds), pp. 59–82. Amsterdam: Elsevier.

Roberts AC, Robbins TW, Everitt BJ (1988) The effects of intradimensional and extradimensional shifts on visual discrimination learning in humans and non-human primates. Q J Exp Psychol 40:321–341.

Rogers RD, Andrews TC, Grasby PM, Brooks DJ, Robbins TW (2000) Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. J Cogn Neurosci 12:142–162.

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. J Neurosci 13:900–913.

Stuss DT, Levine B, Alexander MP, Hong J, Palumbo C, Hamer L, Murphy KJ, Izukawa D (2000) Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes. Neuropsychologia 38:388–402.

Sutton RS (1988) Learning to predict by the method of temporal differences. Machine Learn 3:9–44.

Wilson FAW, Scalaidhe SPO, Goldman-Rakic PS (1993) Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 260:1955–1957.

Appendix — Implementational Details

The model was implemented using the Leabra framework, which is described in detail in O'Reilly and Munakata (O'Reilly and Munakata, 2000) and O'Reilly (O'Reilly 2001), and summarized here. See Table A1 for a listing of parameter values, nearly all of which are at their default settings. These same parameters and equations have been used to simulate over 40 different models in O'Reilly and Munakata (O'Reilly and Munakata 2000), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model. The model can be obtained by emailing the first author at oreilly@psych.colorado.edu.

Pseudocode

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm described in more detail in the subsequent sections fit together.

Outer loop: Iterate over events (trials) within an epoch. For each event:

```
1. Iterate over minus and plus phases of settling for each event.
  (a) At start of settling, for all units:
    i. Initialize all state variables (activation, v_m, etc.).
    ii. Apply external patterns (clamp input in minus, input and output in plus).
  (b) During each cycle of settling, for all non-clamped units:
    i. Compute excitatory net input (ge(t) or hj, equation 3).
    ii. Compute kWTA inhibition for each layer, based on $g_i^\Theta$ (equation 7):
      A. Sort units into two groups based on $g_i^\Theta$ : top $k$ and remaining $k + 1$ to $n$.
      B. If basic, find $k$ and $k + 1$th highest; if avg-based, compute avg of $1 \rightarrow k$ and $k + 1 \rightarrow n$.
      C. Set inhibitory conductance $g_i$ from $g_k^\Theta$ and $g_{k+1}^\Theta$ (equation 6).
    iii. Compute point-neuron activation combining excitatory input and inhibition (equation 1).
  (c) After settling, for all units:
    i. Record final settling activations as either minus or plus phase ($y_j^-$ or $y_j^+$).
2. After both phases update the weights (based on linear current weight values), for all connections:
  (a) Compute error-driven weight changes (equation 9) with soft weight bounding (equation 10).
  (b) Compute Hebbian weight changes from plus-phase activations (equation 8).
  (c) Compute net weight change as weighted sum of error-driven and Hebbian (equation 11).
  (d) Increment the weights according to net weight change.
```

Point Neuron Activation Function

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

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

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t)\bar{g}_c\left(E_c - V_m(t)\right) \tag{1}$$

with four channels (c) corresponding to: e, excitatory input; l, leak current; i, inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a

Table A1
Parameters for the simulation (see equations in text for explanations of parameters)

Parameter	Value	Parameter	Value
E_l	0.15	g_l	0.10
E_i	0.15	g_i	1.0
E_e	1.00	g_e	1.0
V_{rest}	0.15	$\Theta\_$	0.25
τ	0.02	γ	600
k Post Ctx	2†	k Output	1†
k Feat PFC	2†	k Dim PFC	1†
k_{hebb}	0.02	ε	0.01
to AC ε	0.04†		

All are standard default parameter values except for those with a † (most of which have no default because they are intrinsically task-dependent). The faster learning rate (ε) for connections into the AC was important for ensuring rapid learning of reward.

constant $\bar{g}_c$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) to 0:

$$V_m^{\infty} = \frac{g_e \bar{g}_e}{g_e \bar{g}_e + g_l \bar{g}_l + g_i \bar{g}_i} \tag{2}$$

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

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

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

The inhibitory conductance is computed via the kWTA (k-Winners-Take-All) function described in the next section, and leak is a constant.

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

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

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $x > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning mechanisms (e.g. gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0$, $\sigma = 0.005$), which reflects the intrinsic processing noise of biological neurons:

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

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

k-Winners-Take-All Inhibition

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

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

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

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:

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

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

Hebbian and Error-Driven Learning

For learning, Leabra uses a combination of error-driven and Hebbian learning. The error-driven component is the symmetric midpoint version of the GeneRec algorithm (O'Reilly, 1996), which is functionally equivalent to the deterministic Boltzmann machine and contrastive Hebbian learning (CHL). The network settles in two phases: an expectation (minus) phase where the network's actual output is produced, and an outcome (plus) phase where the target output is experienced, and then computes a simple difference of a pre and postsynaptic activation product across these two phases. For Hebbian learning, Leabra uses essentially the same learning rule used in competitive learning or mixtures-of-Gaussians which can be seen as a variant of the Oja normalization. The error-driven and Hebbian learning components are combined additively at each connection to produce a net weight change.

The equation for the Hebbian weight change is:

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

and for error-driven learning using CHL:

$$\Delta_{err} w_{ij} = x_i^+ y_j^+ - x_i^- y_j^- \tag{9}$$

which is subject to a soft-weight bounding to keep within the 0–1 range:

$$\Delta_{sberr} w_{ij} = [\Delta_{err}]_+ (1 - w_{ij}) + [\Delta_{err}]_- w_{ij} \tag{10}$$

The two terms are then combined additively with a normalized mixing constant k_{hebb}:

$$\Delta w_{ij} = \varepsilon [k_{hebb} (\Delta_{hebb}) + (1 - k_{hebb}) (\Delta_{sberr})] \tag{11}$$

Temporal Differences and Adaptive Critic Gating Mechanisms

To implement the temporal differences (TD) algorithm in Leabra, the adaptive critic (AC) unit in the VTA layer has plus–minus phase states that correspond to the expected reward at the previous time step (minus) and the current time step (plus). The difference between these two states is the TD error δ, which is essentially equivalent to the more standard kinds of error signals computed by the error-driven learning component of Leabra, except that it represents an error of prediction over time, instead of an instantaneous error in the network output.

The AC–PFC relationship is formalized in the model with the following equations for the gating (multiplicative scaling) terms s_{in} (the weight scaling of the PFC inputs) and s_{maint} (the weight-scaling of the PFC self-maintenance connections):

$$s_{\mathrm{in}} = b_{\mathrm{in}} + \delta + \nu \quad (12)$$

$$s_{\mathrm{maint}} = b_{\mathrm{maint}} + \delta + \nu \quad (13)$$

where δ is the change in AC activation (TD error), and ν is a Gaussian random noise value that allows for random trial-and-error exploration ($\mu = 0$, $\sigma = 0.2$). The base-level parameters b_{in} and b_{maint} determine the basic level of each weight-scaling (gain) parameter, and are set to 0 and 1, respectively. Both of the weight-scaling terms are bounded between 0 and 1. These differences in input and maintenance connections could result from different dopamine receptor affinities, and have the effect that the inputs tend to weakly impact the PFC units except during a positive δ, while the maintenance connections are relatively strong except during a negative δ.