Prefrontal cortex and flexible cognitive control: Rules without symbols

Nicolas P. Rougier*†, David C. Noelle‡, Todd S. Braver§, Jonathan D. Cohen¶, and Randall C. O'Reilly*‖

*Department of Psychology, University of Colorado, 345 UCB, Boulder, CO 80309; †Institut National de Recherche en Informatique et en Automatique Lorraine, Campus Scientifique, B.P. 239, F-54506 Vandoeuvre-Lès-Nancy Cedex, France; ‡Department of Electrical Engineering and Computer Science, Vanderbilt University, Vu Station B 351679, Nashville, TN 37235; §Department of Psychology, Washington University, Campus Box 1125, St. Louis, MO 63130-4899; and ¶Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544

Communicated by James L. McClelland, Carnegie Mellon University, Pittsburgh, PA, March 29, 2005 (received for review February 3, 2005)

Human cognitive control is uniquely flexible and has been shown to depend on prefrontal cortex (PFC). But exactly how the biological mechanisms of the PFC support flexible cognitive control remains a profound mystery. Existing theoretical models have posited powerful task-specific PFC representations, but not how these develop. We show how this can occur when a set of PFC-specific neural mechanisms interact with breadth of experience to self organize abstract rule-like PFC representations that support flexible generalization in novel tasks. The same model is shown to apply to benchmark PFC tasks (Stroop and Wisconsin card sorting), accurately simulating the behavior of neurologically intact and frontally damaged people.

generalization | abstraction | adaptive gating

A fundamental human cognitive faculty is the capacity for cognitive control: the ability to behave in accord with rules, goals, or intentions, even when this runs counter to reflexive or otherwise highly compelling competing responses (e.g., the ability to keep typing rather than scratch a mosquito bite). A hallmark of cognitive control in humans is its remarkable flexibility: we can perform novel tasks with very little additional experience (e.g., playing a card game for the first time by observing the play or hearing the rules described). This ability appears to depend on the prefrontal cortex (PFC) (1–5) and in particular on abstract rule-like representations localized to this brain area (6–8). However, this capacity emerges only slowly over a protracted period through late adolescence, closely tracking the development of the PFC (9–11). At the psychological level, flexible cognitive control has been modeled abstractly in terms of symbol processing computations that support arbitrary variable binding (12). However, it remains unclear whether or how such models correspond to the increasingly rich body of knowledge about the neural mechanisms underlying cognitive control and in particular the functioning of the PFC. At the biological level, a number of neural models have proposed that cognitive control relies on the active maintenance of abstract rule-like representations in PFC that guide processing in posterior cortex (13–17). However, none of these existing frameworks have explained how such representations might develop, and why this development should take so long; indeed, most models rely on hand-coded representations designed explicitly for solving a specific set of tasks. Thus, a major challenge to theories of the neural bases of cognitive control remains unanswered: how it can be explained in terms of self-organizing mechanisms that develop on their own, over time, without recourse to unexplained sources of influence or intelligence (i.e., a "homunculus") (18).

Here, we present a computational model that provides an explanation for the development of cognitive flexibility. This model shows how neurobiological mechanisms specific to the PFC result in the self organization of abstract rule-like PFC representations that support flexible cognitive control. These representations develop through experience on a basic set of sensory-motor tasks via synaptic learning mechanisms. Both the development of these representations and the flexibility they support required a broad range of experience across multiple tasks. Thus, this model describes a biologically based alternative to abstract symbol processing models of cognitive flexibility that illustrates how cognitive flexibility can arise from an interaction between nature (PFC-specific neurobiological mechanisms) and nurture (breadth of experience). Our model builds on extensive neurobiological and theoretical work indicating that PFC exhibits the following properties (see supporting information, which is published on the PNAS web site, for details of the implementation):

(*i*) Active maintenance of patterns of neural activity over time and against interference from distracting inputs, so that currently relevant information can be held in working memory (1–3). Both recurrent excitatory connectivity that sustains active patterns of PFC neural activity and intrinsic bistability of PFC neurons have been shown to support active maintenance (19, 20), and both of these mechanisms are included in our model.

(*ii*) Adaptive updating of these PFC activity patterns by dynamically switching between active maintenance and rapid updating of new representations (16, 17, 21, 22). This updating function is implemented by an adaptive gating mechanism based on the circuits and physiology of the basal ganglia and the midbrain dopaminergic ventral tegmental area (VTA), which project extensively to the PFC (16, 17, 23, 24). This gating mechanism leverages the close formal relationship between VTA dopamine firing and reinforcement learning based on expected rewards (25). Specifically, the gating system stabilizes and destabilizes active maintenance in the PFC and is itself driven by differences in expected and received rewards. When the gating system receives an unexpected reward, the corresponding dopamine spike stabilizes active representations in the PFC by activating intrinsic maintenance currents; when it does not get an expected reward, it destabilizes the PFC to allow a new activation pattern to emerge. This allows PFC representations to rapidly update to reflect changing task contingencies. We have also explored the idea that the basal ganglia provide a direct gating input to the PFC (23), which is trained by similar dopamine-based mechanisms but can provide reliable gating in the absence of dopamine signals and also a more selective updating signal.

(*iii*) PFC modulation of processing in other cortical areas (e.g., in posterior cortex) responsible for task execution (3, 13), supported by extensive interconnectivity with these other cortical areas (2).

We present the results of two simulation experiments using the model. The first shows that the model's mechanisms are sufficient to support the development of rule-like task representations, and that these representations support generalization of task performance to novel environments. The second shows that the model accurately simulates detailed patterns of behavior from neurolog-

Abbreviations: PFC, prefrontal cortex; WCST, Wisconsin Card Sort Task.

‖To whom correspondence should be addressed. E-mail: oreilly@psych.colorado.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0502455102

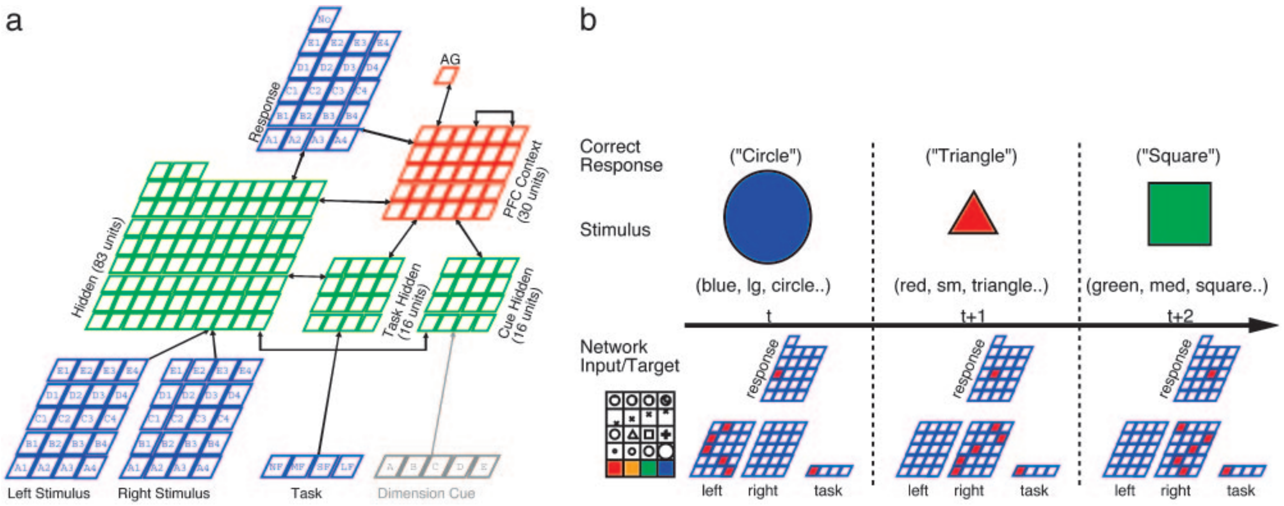

Fig. 1. Model and example stimuli. (*a*) The model with the complete PFC system. Stimuli are presented in two possible locations (left, right). Rows represent different stimulus dimensions (e.g., color, size, shape, etc., labeled A–E for simplicity), and columns represent different features (red, orange green, and blue; small, medium, etc., numbered 1–4). Other inputs include a task input indicating current task to perform (NF, name feature; MF, match feature; SF, smaller feature; LF, larger feature), and, for the ''instructed'' condition (used to control for lack of maintenance in non-PFC networks), a cue to the currently relevant dimension. Output responses are generated over the response layer, which has units for the different stimulus features, plus a ''No'' unit to signal nonmatch in the matching task. The hidden layers represent posterior cortical pathways associated with different types of inputs (e.g., visual and verbal). The AG unit is the adaptive gating unit, providing a temporal differences (TD) based dynamic gating signal to the PFC context layer. The weights into the AG unit learn via the TD mechanism, whereas all other weights learn using the Leabra algorithm that combines standard Hebbian and error-driven learning mechanisms, together with k-winners-take-all inhibitory competition within layers and point-neuron activation dynamics (26) (also see supporting information). (*b*) Example stimuli and correct responses for one of the tasks (NF) across three trials where the current rule is to focus on the Shape dimension (the same rule was blocked over 200 trials to allow networks plenty of time to adapt to each rule). The corresponding input and target patterns for the network are shown below each trial, with the unit meanings given by the legend in the lower left. The network must maintain the current dimension rule to perform correctly.

ically intact and frontally damaged people on benchmark tasks of cognitive control.

Methods

We tested a model implementing the three sets of PFC-specific mechanisms described above (Fig. 1*a*), as well as versions of it lacking these mechanisms by varying degree. These models were trained either on two (Task Pairs condition) or four tasks (All Tasks condition), to test the effects of restricted vs. broad training experience, respectively. The tasks were designed to simulate simple processing of multidimensional stimuli (e.g., varying along dimensions such as size, shape, color, etc.) and active maintenance. Critically, we constructed these tasks so they all shared a common requirement: only one stimulus dimension was relevant at a given time. For example, one task involved naming a stimulus feature value along a given dimension (e.g., if the stimulus was a blue large circular object, and the relevant dimension was shape, then the correct response was to activate the "circle" output unit; Fig. 1*b*). Other tasks included matching features of two stimuli (if they matched along the relevant dimension, the correct output was the name of the shared feature; otherwise, the "No Match" unit should be activated) or comparing their relative ordinal values (i.e., output the name of the larger/smaller feature within the relevant dimension).

Thus, knowing the relevant dimension was a critical rule in each task, uniquely determining the mapping from stimulus to response. Because all of the tasks shared this requirement, attention to a single dimension, we predicted that during training, the PFC would develop abstract representations of these dimensions (i.e., learn the relevant set of rules), and that this would allow it to generalize its performance to novel stimuli in each task. To allow the current rule to be discovered solely by trial-and-error learning (even in networks without a PFC, which adapted relatively slowly to task rule changes), we kept the relevant dimension the same over blocks of trials (a variety of strategies for blocking task and dimension information were explored without substantial differences in results, as described in supporting information; the basic case was task switching every block of 25 trials, with dimension switching after two iterations through all of the tasks). These conditions were designed to simulate simple forms of real-world learning experience that humans encounter during development (e.g., in playing with blocks, a sustained focus on the shapes of these objects is necessary to construct desired structures). Furthermore, we also included the ability to provide explicit task instructions to the models by means of a dimension cue input, to provide as generous a test as possible of models lacking the ability to maintain task-relevant information internally (see supporting information for more details and effects of parametric variations).

To enable generalization testing, the model saw only a subset of the feature values along each dimension for a given task and a relatively small fraction ($\approx$30%) of all possible stimuli (i.e., combinations of features across dimensions). A given training run consisted of 100 epochs of 2,000 trials per epoch; it took the networks only $\approx$10 epochs to achieve near-perfect performance on the training items, but we measured crosstask generalization performance every five epochs throughout the duration to find the best generalization for each network, unconfounded by any differences in architecture or in the raw amount of exposure to features across different training scenarios. Generalization testing measured the network's ability to respond to stimuli it had not seen in that task.

We trained and tested different network configurations to test the contribution made by constituent mechanisms to learning and performance. All network configurations had the same total number of processing units, to control for the effects of overall computing resources. The only differences among configurations were the patterns of connectivity and the presence or absence of the adaptive gating mechanism. The various configurations are described in Fig. 3. These ranged from a simple feedforward network with 145 hidden units (equaling the number of hidden plus PFC units in the full PFC model) to the complete model, including full recurrent connectivity within the PFC and an adaptive gating mechanism. For all networks, we ran 10 different random initial

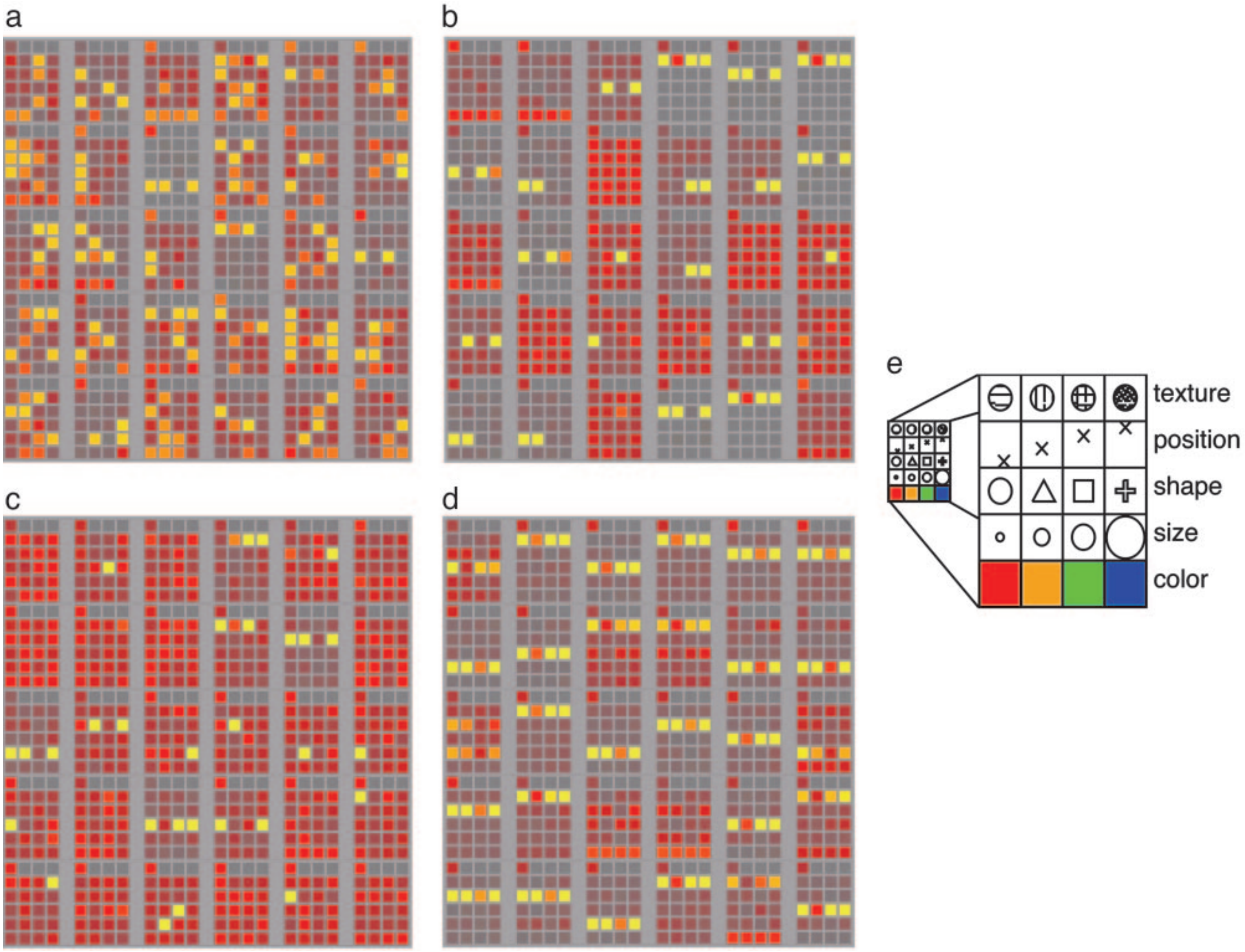

Fig. 2. Representations (synaptic weights) that developed in four different network configurations. (*a*) Posterior cortex only (no PFC) trained on all tasks. (*b*) PFC without the adaptive gating mechanism (all tasks). (*c*) Full PFC trained only on task pairs (name feature and match feature in this case). (*d*) Full PFC (all tasks). Each image shows the weights from the hidden units (*a*) or PFC (*b*–*d*) to the response layer. Larger squares correspond to units (all 30 in the PFC and a random and representative subset of 30 from the 145 hidden units in the posterior model), and the smaller squares within designate the strength of the connection (lighter = stronger) from that unit to each of the units in the response layer. Note that each row designates connections to response units representing features in the same stimulus dimension (as illustrated in *e* and Fig. 1). It is evident, therefore, that each of the PFC units in the full model (*d*) represents a single dimension and, conversely, that each dimension is represented by a distinct subset of PFC units. This pattern is less evident to almost entirely absent in the other network configurations (see text for additional analyses).

networks to generate statistics, and error bars in Figs. 3 and 4 reflect the standard error over these runs.

The model was implemented in the Leabra algorithm, which includes error-driven and associative (Hebbian) learning mechanisms, k-winners-take-all inhibitory competition within layers, and point-neuron ion-channel-based neural dynamics with bidirectional excitatory connectivity. Leabra integrates the most widely used neural modeling principles developed by a variety of researchers into one unified framework, which has been used to simulate >40 different cognitive models from perception and attention to learning, memory, language, and higher-level cognition (26), plus many more published simulations in other papers. In keeping with the goal of using the same set of mechanisms and parameters across a wide range of models, default parameters and mechanisms were used in this model. The details of these standard mechanisms and the PFC-specific mechanisms in our model are described in ref. 24 and supporting information.

Results

Representations and Generalization. Our primary finding was that, over the course of training on these tasks, the PFC layer in the full model developed synaptic weights and associated patterns of activity that encoded abstract rule-like representations of the relevant stimulus dimensions (Fig. 2*d*). That is, each PFC unit came to represent a single dimension and all features in that dimension. More precisely, these representations collectively formed a basis set of orthogonal vectors that spanned the space of task-relevant stimuli, and that were aligned with the dimensions along which features had to be distinguished for task performance. More generally, we can characterize rule-like representations as encoding and producing a common abstract pattern of behavior over a broad class of specific situations. These representations were only partially apparent in the configuration having a PFC but lacking an adaptive gating mechanism (Fig. 2*b*), as well as the full model trained only on task pairs (Fig. 2*c*), and were essentially absent from the model entirely lacking a PFC (Fig. 2*a*). These models tended to memorize specific combinations of stimulus features and responses rather than develop abstract representations of feature dimensions that could serve as more general rules. Additional principal components analysis supported this visual interpretation of the weights, showing that the non-PFC networks do not simply have a low-dimensional "rotated" representation of the dimensions (e.g., the posterior cortex model had 8 eigenvalues >1 and a smooth continuum down to a minimum of 0.4, which is still relatively large). As noted in

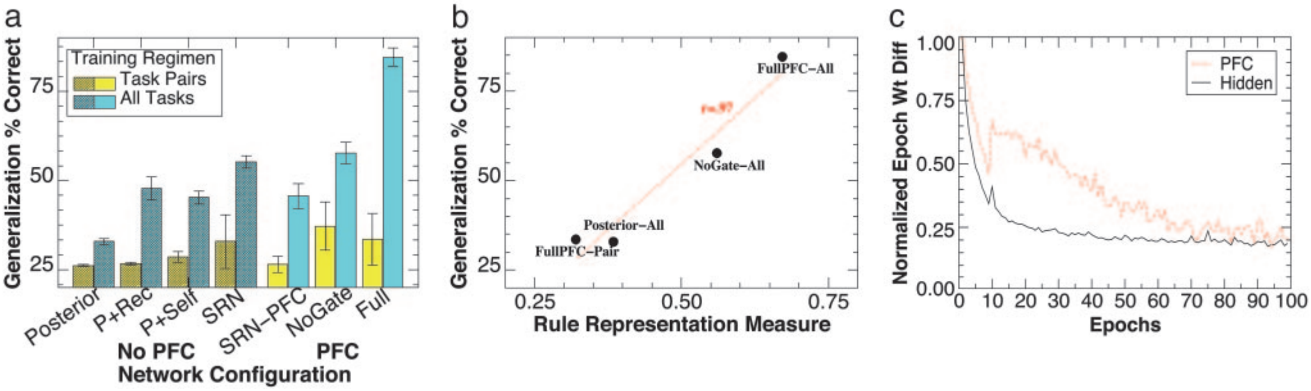

Fig. 3. Generalization and learning results. (*a*) Crosstask generalization results (% correct on task-novel stimuli) for the full PFC network and a variety of control networks, with either only two tasks (Task Pairs) or all four tasks (All Tasks) used during training ($n = 10$ for each network, error bars are standard errors). Overall, the full PFC model generalizes substantially better than the other models, and this interacts with the level of training such that performance on the All Tasks condition is substantially better than the Task Pairs condition (with no differences in numbers of training trials or training stimuli). With one feature left out of training for each of four dimensions, training represented only 31.6% (324) of the total possible stimulus inputs (1,024); the ≈85% generalization performance on the remaining test items therefore represents good productive abilities. The other networks are: Posterior, a single large hidden unit layer between inputs and response, a simple model of posterior cortex without any special active maintenance abilities; P + Rec, posterior + full recurrent connectivity among hidden units, allows hidden layer to maintain information over time via attractor dynamics; P + Self, posterior + self-recurrent connections from hidden units to themselves, allows individual units to maintain activations over time; SRN, simple recurrent network, with a context layer that is a copy of the hidden layer on the prior step, a widely used form of temporal maintenance; SRN-PFC, an SRN context layer applied to the PFC layer in the full model (identical to the full PFC model except for this difference), tests for role of separated hidden layers; NoGate, the full PFC model without the AG adaptive gating unit. (*b*) The correlation of generalization performance with the extent to which the units distinctly and orthogonally encode stimulus dimensions for the networks shown in Fig. 2. This was computed by comparing each unit's pattern of weights to the set of five orthogonal, complete dimensional target patterns (i.e., the A dimension target pattern has a 1 for each A feature, and 0s for the features in all other dimensions, etc.). A numeric value between 0 and 1, where 1 represents a completely orthogonal and complete dimensional representation was computed for unit *i* as: $d_i = \max_k |w_i \cdot t_k| / \Sigma_k |w_i \cdot t_k|$; where t_k is the dimensional target pattern *k*, and w_i is the weight vector for unit *i*, and $|w_i \cdot t_k|$ represents the normalized dot product of the two vectors (i.e., the cosine). This value was then averaged across all units in the layer and then correlated with that network's generalization performance. (*c*) Relative stability of PFC and hidden layer (posterior cortex) in the model, as indexed by Euclidean distance between weight states at the end of subsequent epochs (epoch = 2,000 trials). The PFC takes longer to stabilize (i.e., exhibits greater levels of weight change across epochs) than the posterior cortex. For PFC, within-PFC recurrent weights were used. For Hidden, weights from stimulus input to Hidden were used. Both sets of weights are an equivalent distance from error signals at the output layer. The learning rate is reduced at 10 epochs, producing a blip at that point.

Methods, the total number of training trials and stimulus inputs was equated across simulation conditions, so that the increased breadth of experience in the All Tasks condition was solely from exposure to more task contexts. Furthermore, models were trained well beyond convergence, so differences in overall learning rate are not a factor.

The abstract rule-like representations that developed in the full PFC model supported task performance by providing top-down excitatory support for the relevant stimulus dimension in the rest of the network. The adaptive gating system learned to update the PFC layer activity when the relevant stimulus dimension (i.e., task rule) changed (due to rapid error-based destabilization of PFC activations), and the PFC actively maintained this rule while it remained in effect. In models without these active maintenance and updating mechanisms, synaptic learning mechanisms shifted the network's processing to the relevant stimulus dimension, but these changes were necessarily slower than the rapid shifts that can be achieved by dynamic updating of activation states in PFC (26). This difference accounts for the increased levels of perseveration observed with PFC damage in the Wisconsin Card Sort Task (WCST) and other tasks, as has been demonstrated in several existing models (14, 15, 24) and as we report for our model below.

We hypothesized that the abstract rule-like representations that developed in the full PFC model should support more flexible cognitive control in this model relative to the others. We tested this idea by comparing the ability of each network to generalize its performance across the different tasks. Each network was trained on a subset of stimuli in each task and then tested on stimuli that it had not previously seen in that task. We theorized that the abstract dimensional representations in the PFC would be able to guide processing for the task–novel test stimuli in a similar manner as the trained stimuli. Indeed, only the Full PFC model exhibited substantial generalization, achieving 85% accuracy (i.e., only one-third as many errors as other networks) on stimuli for which it had no prior same-task experience (Fig. 3*a*). However, this was the case only for the All Tasks regimen; training on pairs of tasks resulted in more than four times as many generalization errors. This indicates that breadth of experience was critical for exploiting the mechanisms present in the PFC, just as we had earlier observed in the development of the abstract rule-like PFC representations. Indeed, Fig. 3*b* shows that, as we hypothesized, the degree to which different networks developed abstract dimensional representations was strongly correlated with the network's generalization performance ($r = 0.97$).

There is a clear mechanistic explanation for why the combination of rapid updating and sustained active maintenance of task rule representations in the full PFC model (which depends on the adaptive gating mechanism) was critical for the formation of abstract rule-like representations during training. Within a block of trials with the same relevant dimension, the specific features within that dimension varied, but a constant PFC activity pattern was maintained due to the gating mechanism. This caused these PFC representations, which initially had random connections, to begin to encode all of the varying features within a dimension, resulting in an abstract dimensional representation. In contrast, other networks tended to activate new representations for each new stimulus (as the specific features changed) and thus were unable to form the dimensional abstraction across features. Interestingly, the dimensional alignment of PFC representations was greater for the All Tasks than the Task Pairs condition. This is because the pressure to use the same PFC representations across all tasks increased with the number of tasks: with only two tasks, it was possible for the network

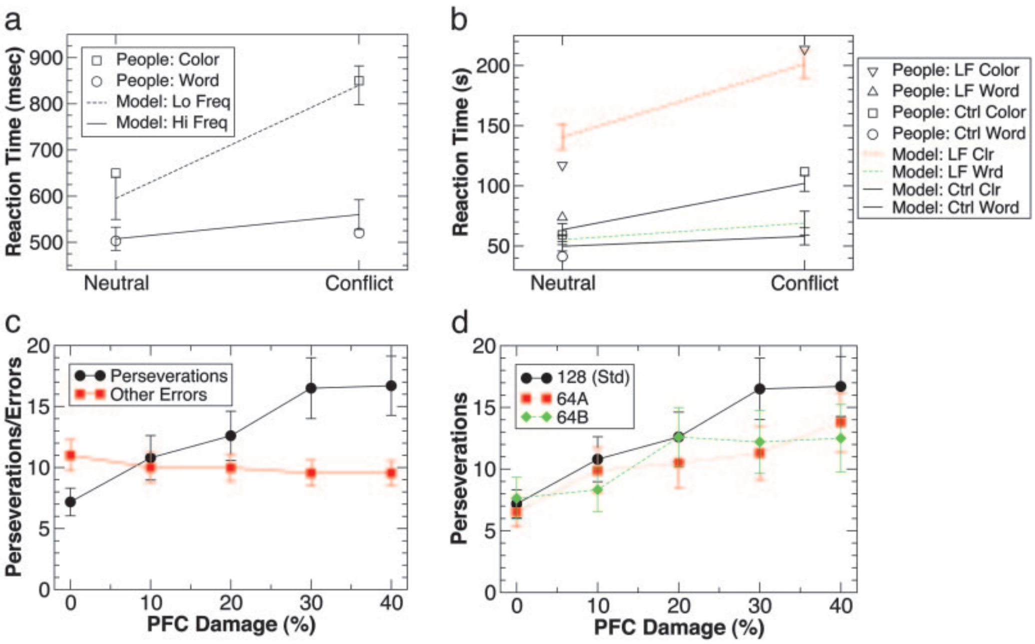

Fig. 4. Neuropsychological task results. (*a*) Performance of the full PFC network on a simulated Stroop task, demonstrating the classic pattern of conflict effects on the subordinate task of color naming with unaffected performance on the dominant word reading task (human data from ref. 31). This was simulated by training one dimension (*a*) with one-fourth the frequency of the others, making it weaker. In the neutral condition, a single feature was active, whereas the conflict condition had two features present and the dimension cue input specified that was to be named. Reaction time (RT) was measured as the number of cycles to activate a feature in the response layer >0.75 (multiplied by 35 to match human RT in msec). (*b*) Stroop performance for a 30% lesion (removal) of PFC units in the model (posttraining), compared with data from ref. 30 on patients with left frontal (LF) lesions (six of eight include dorsolateral PFC) and matched controls (Ctrl) (data in seconds to complete a block of trials; model cycles were transformed as RT = cycles × 5.5–30 to fit this scale; the Conflict Word reading conditions were not run on the human subjects). The main effect of damage is an overall slowing of color naming, consistent with the notion that the PFC provides top-down support to this weaker pathway via abstract dimensional representations. (*c*) Performance in a simulated WCST task, demonstrating the classic pattern of increasing perseveration with increased PFC damage (% of units removed, posttraining). Perseverations = number of sequential productions of feature names corresponding to the previously relevant dimension after a switch. Clearly, the simulated PFC is critical for rapid flexible switching. (*d*) WCST results (perseverations) for the three different training conditions used by ref. 28 (128 is the standard case plotted before, whereas 64A involves providing instructions about the relevant dimensions along which cards could be sorted, and 64B has explicit instruction when the rule changes; see supporting information for details). n = 10 networks; error bars = standard error for all graphs.

to use different PFC representations for different tasks, but this strategy becomes less and less efficient as the number of tasks increases. The adaptive gating mechanism also caused the PFC representations to focus on single dimensions, instead of encoding features across multiple dimensions, because the gating mechanism caused all active PFC units to be inhibited upon a dimension switch, discouraging persistent activation across multiple dimensions. Thus, overall, the adaptive gating mechanism plays a critical role in shaping the PFC representations.

Our model makes the additional prediction that PFC representations should stabilize later in development (training) than those in posterior areas, because it is necessary for representations in posterior systems to stabilize before the PFC can extract the dimensions of these representations relevant to task performance. We tested this by measuring the average magnitude of weight changes from projections into the main hidden (posterior cortex) layer and in the PFC layer. The hidden layer stabilized within 20 epochs (one epoch is 2,000 trials), whereas the PFC did not stabilize until 70 epochs (Fig. 3*c*). This slower development of PFC representations, together with the breadth of training required, is consistent with the protracted developmental course of the human PFC (extending into late adolescence), which allows a broad range of experience to shape PFC representations (9–11).

Neuropsychological Tasks. We next explored whether the rule-like PFC representations learned by our model can produce appropriate patterns of performance in tasks specifically associated with prefrontal function. To do so, we used the full PFC model trained in the All Tasks condition to perform simulations of the Stroop task and the WCST, two tasks that have been used widely as benchmarks of prefrontal function (27–30). Converging evidence from a variety of sources suggests that the kinds of dimensional stimulus representations found in our model are localized in dorsolateral areas of PFC (DLPFC) in humans (see supporting information for more discussion). Accordingly, we focused on DLPFC lesion data in both of these tasks.

In the Stroop task, participants are presented with color words printed in various colors and are asked to either read the word or name the color in which it is printed. Due to greater familiarity with word reading, it is relatively faster than color naming, and an incongruent word (e.g., "green" displayed in red) interferes with color naming (saying "red"), whereas word reading is relatively unaffected. To simulate these asymmetries of experience in our model, one of the stimulus dimensions was trained less (25% as much) than the other four dimensions, with all other factors unchanged from the first study. The model captures the characteristic effects seen in human Stroop performance (Fig. 4*a*). These results replicate previous modeling work showing that top-down excitation from PFC representations of the dimensions that define each task (colors vs. words) can partially compensate for the differences in relative strength of the relevant posterior pathways (13, 26). However, unlike these earlier models, PFC representations in our model developed through learning. Furthermore, Fig. 4*b* shows that simulated lesions to the model's PFC layer (30% unit removal, post training) replicate the color-naming impairments observed from PFC lesions (predominantly dorsolateral areas of PFC) in human patients (30), consistent with the observation that this PFC area supports abstract color dimension representations (29).

In the WCST task, participants are provided with a deck of cards bearing multidimensional stimuli that vary in shape, size, color, and number. These must be sorted according to a particular dimension (rule), which must be discovered from trial-and-error feedback. The rule switches without warning after the participant makes a criterion number of correct responses in sequence (e.g., ref. 8). Patients with frontal damage typically are able to discover the first rule without difficulty, but after a switch, they perseverate in sorting according to the previous rule. This and other similar findings have led many authors to conclude that PFC plays a critical role in the cognitive flexibility required to switch "mental set" from one rule to another (4). In our model, we used the feature-naming task to simulate the WCST: a stimulus is presented, and the feature value in the relevant dimension must be output. The relevant dimension is discovered via trial-and-error learning and switches after eight correct responses in a row. Fig. 4*c* shows that increasing amounts of PFC damage (unit removal and post training) produce a disproportionate increase in perseverative responding relative to other types of errors [consistent with earlier modeling studies with manually imposed PFC representations (14, 15)]. Furthermore, the model successfully reproduced the modest effects on perseveration (Fig. 4*d*) that were observed with various levels of additional instruction provided by Stuss *et al.* (28).

Discussion

The findings reported here provide insight into how the capacity for flexible cognitive control can develop without invoking unexplained forms of intelligence (i.e., a "homunculus"). Our model shows how specialized neural mechanisms that support adaptive updating of active maintenance interact with breadth of learning experience to produce abstract rule-like representations in the PFC. These PFC representations produced significantly higher levels of generalization across tasks by guiding stimulus processing according to abstract dimensions that apply across both familiar and task-novel stimuli. This crosstask generalization is an important measure of cognitive flexibility. Thus, the model illustrates how nature and nurture can interact to produce human cognitive abilities. It explains in explicit mechanistic terms why rule-like representations are predominantly found in the PFC (6–8), and why cognitive flexibility, dependent upon the biological substrate of the PFC, takes a long time to develop, extending into late adolescence (9–11).

Although we found that abstract rule-like PFC representations supported good generalization in the fully regular domains that we explored here, we do not claim that these representations are universally beneficial. In particular, it is unlikely that such discrete abstract representations are as useful in task domains characterized by more graded knowledge structures, where distributed representations may perform better (e.g., perceptual categorization, face recognition, etc.). Thus, there may be a tradeoff between PFC and posterior cortical forms of representation, in which each is better suited for different types of tasks. This is consistent with data showing that the posterior cortex may be better at learning complex similarity-based categories, whereas PFC can more quickly acquire simple rule-based categories (32). More work is needed to explore these potential tradeoffs, for example, in richer more complex domains such as language, wherein our model may provide a productive middle ground between the neural network and symbolic modeling perspectives in the longstanding "rules and regularities in language processing" debates (33).

The model illustrates another critical factor that contributes to flexibility of control: the use of patterns of activity rather than changes in synaptic weights as a means of exerting control over processing (26, 34). We showed that PFC representations in our model developed slowly over many trials of synaptic modification. However, once these were learned, adaptive behavior in novel circumstances was mediated by a search for the appropriate pattern of activity (using simple principles of reinforcement learning), rather than the need to learn a new set of connection strengths. This may clarify the mechanisms underlying the adaptive coding hypothesis (5), which holds that PFC dynamically reconfigures itself for the task at hand. Importantly, this activation-based processing differs fundamentally from the arbitrary variable binding mechanisms of traditional symbolic models (12), where the meaning of the underlying representations (symbols) can be arbitrarily bound to novel inputs to achieve flexible performance. Thus, the representations in our model produce rule-like behavior without implementing biologically problematic symbolic processing computations.

The tasks used in our simulations were relatively simple, with the common requirement that the network selectively process one dimension of information. Nevertheless, the principles developed here are likely to apply in more realistic task domains, where the relevant rules may be more complex. These complex rule representations must also be maintained over a sequence of behaviors operating on specific stimuli (e.g., rules of a card game applied over different rounds of play), to guide behavior in a more systematic fashion. Thus, the learning mechanisms in our model, which form abstract rule-like representations by integrating over trials of processing specific instances of the rule, should also apply in these cases.

Finally, although our model provides an important step toward understanding the neurobiological mechanisms underlying flexible human cognitive control, it captures only a subset of such mechanisms. An understanding of how PFC representations can be dynamically recombined and can interact with other systems (such as those supporting episodic memory, language function, and affect) will be equally important in developing a full understanding of how cognitive control is implemented in the brain.

We thank Carlos Brody, Tim Curran, Michael Frank, Tom Hazy, Dave Jilk, Ken Norman, Yuko Munakata, Alex Petrov, and members of the Computational Cognitive Neuroscience lab for helpful comments. This work was supported by Office of Naval Research Grants N00014-00-1-0246 and N00014-03-1-0428 and National Institutes of Health Grants MH64445 and MH069597.

1. Goldman-Rakic, P. S. (1987) *Handb. Physiol.* **5,** 373–417.
2. Fuster, J. M. (1997) *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe* (Lippincott–Raven, New York), 3rd Ed.
3. Miller, E. K. & Cohen, J. D. (2001) *Annu. Rev. Neurosci.* **24,** 167–202.
4. Shallice, T. (1988) *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, New York).
5. Duncan, J. (2001) *Nat. Rev. Neurosci.* **2,** 820–829.
6. White, I. M. & Wise, S. P. (1999) *Exp. Brain Res.* **126,** 315–335.
7. Wallis, J. D., Anderson, K. C. & Miller, E. K. (2001) *Nature* **411,** 953–956.
8. Sakai, K. & Passingham, R. E. (2003) *Nat. Neurosci.* **6,** 75–81.
9. Diamond, A. & Goldman-Rakic, P. S. (1989) *Exp. Brain Res.* **74,** 24–40.
10. Huttenlocher, P. R. (1990) *Neuropsychologia* **28,** 517–527.
11. Morton, J. B. & Munakata, Y. (2002) *Dev. Sci.* **5,** 435–440.
12. Newell, A. & Simon, H. A. (1972) *Human Problem Solving* (Prentice–Hall, Englewood Cliffs, NJ).
13. Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990) *Psychol. Rev.* **97,** 332–361.
14. Dehaene, S. & Changeux, J. P. (1991) *Cereb. Cortex* **1,** 62–79.
15. O'Reilly, R. C., Noelle, D, Braver, T. S. & Cohen, J. D. (2002) *Cereb. Cortex* **12,** 246–257.
16. Braver, T. S. & Cohen, J. D. (2000) in *Control of Cognitive Processes: Attention and Performance*, eds. Monsell, S. & Driver, J. (MIT Press, Cambridge, MA), XVIII Ed., pp. 713–737.
17. O'Reilly, R. C., Braver, T. S. & Cohen, J. D. (1999) in *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, eds. Miyake, A. & Shah, P. (Cambridge Univ. Press, New York), pp. 375–411.
18. Monsell, S. (1996) in *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*, ed. Bruce, V. (Psychology Press, Hove, U.K.), pp. 93–148.
19. Fellous, J. M., Wang, X. J. & Lisman, J. E. (1998) *Nat. Neurosci.* **1,** 273–275.
20. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. (2000) *J. Neurophysiol.* **83,** 1733–1750.
21. Cohen, J. D., Braver, T. S. & O'Reilly, R. C. (1996) *Philos. Trans. R. Soc. London B* **351,** 1515–1527.
22. Hochreiter, S. & Schmidhuber, J. (1997) *Neural Comput.* **9,** 1735–1780.
23. Frank, M. J., Loughry, B. & O'Reilly, R. C. (2001) *Cognit. Affect. Behav. Neurosci.* **1,** 137–160.
24. Rougier, N. P. & O'Reilly, R. C. (2002) *Cognit. Sci.* **26,** 503–520.
25. Montague, P. R., Dayan, P. & Sejnowski, T. J. (1996) *J. Neurosci.* **16,** 1936–1947.
26. O'Reilly, R. C. & Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, Cambridge, MA).
27. Weinberger, D. R., Berman, K. F. & Daniel, D. G. (1991) in *Frontal Lobe Function and Dysfunction*, eds. Levin, H. S., Eisenberg, H. M. & Benton, A. L. (Oxford Univ. Press, New York), pp. 276–285.
28. Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., Murphy, K. J. & Izukawa, D. (2000) *Neuropsychologia* **38,** 388–402.
29. MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. (2000) *Science* **288,** 1835–1838.
30. Stuss, D. T., Floden, D., Alexander, M. P., Levine, B. & Katz, D. (2001) *Neuropsychologia* **39,** 771–786.
31. Dunbar, K. & MacLeod, C. M. (1984) *J. Exp. Psychol.* **10,** 622–639.
32. Smith, E. E., Patalano, A. L. & Jonides, J. (1998) *Cognition* **65,** 167–196.
33. McClelland, J. L. & Patterson, K. (2002) *Trends Cognit. Sci.* **6,** 465–472.
34. Munakata, Y. (1998) *Dev. Sci.* **1,** 161–184.