Prefrontal Cortex and Flexible Cognitive Control: Rules Without Symbols

Nicolas P. Rougier, David C. Noelle, Todd S. Braver, John D. Cohen, Randall C. O'Reilly

Prefrontal Cortex and Flexible Cognitive Control: Rules Without Symbols

Nicolas P. Rougier[1,2], David C. Noelle[3], Todd S. Braver[4],
Jonathan D. Cohen[5†], Randall C. O'Reilly[1†]

[1]Department of Psychology, University of Colorado Boulder
[2]INRIA Lorraine, France
[3]Department of Electrical Engineering and Computer Science, Vanderbilt University
[4]Department of Psychology, Washington University
[5]Department of Psychology, Princeton University

[†]To whom correspondence should be addressed; E-mail: jdc@princeton.edu, oreilly@psych.colorado.edu

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.

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 novel 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 only emerges 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 can it 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 a novel 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 that 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 (*19*) for details of the implementation):

1. 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 (*20,21*), and both of these mechanisms are included in our model.

2. Adaptive updating of these PFC activity patterns by dynamically switching between active

maintenance and rapid updating of new representations (*16,17,22,23*) 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, 24, 25*). This gating mechanism leverages the close formal relationship between VTA dopamine firing and reinforcement learning based on expected rewards (*26*). 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.

3. 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 neurologically intact and frontally-damaged people on benchmark tasks of cognitive control.

Simulation Study 1

We tested a model implementing the three sets of PFC-specific mechanisms described above (Figure 1a), as well as versions of it lacking these mechanisms by varying degree. These models were trained either on two tasks (Task Pairs condition) or four tasks (All Tasks condition), to test the effects of restricted versus 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 that 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

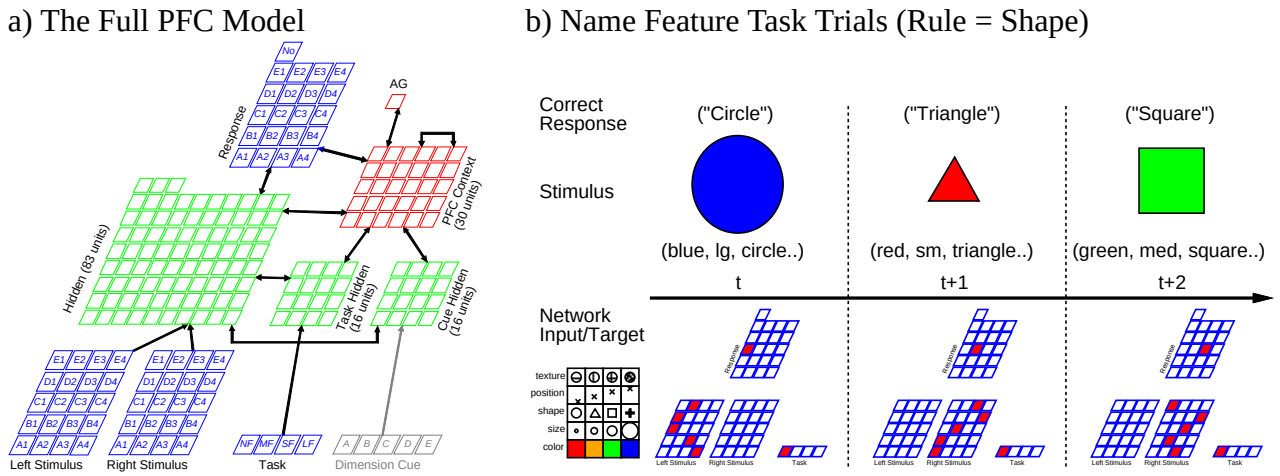

Figure 1: **a)** The model with the complete PFC system. Stimuli are presented in two possible locations (left, right). Rows represent different stimulus dimensions (e.g., color, size, shape, etc., labeled A-E for simplicity), and columns represent different features (red, orange green, blue; small, medium, etc., numbered 1-4). Other inputs include a task input indicating current task to perform (NF = name feature, MF = match features, SF = smaller feature, LF = larger feature), and, for the "instructed" condition (used to control for lack of maintenance in non-PFC networks), a cue as 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 non-match in the matching task. The hidden layers represent posterior cortical pathways associated with different types of inputs (e.g., visual, 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, while 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 (*19, 27, 28*). **b)** Example stimuli and correct responses for one of the tasks (Name Feature; 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.

response was "circle"; Figure 1b). Other tasks included matching features of two stimuli, or their relative ordinal values, along a given stimulus 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 that adapted relatively slowly to task rule changes), we kept the relevant dimension the same over blocks of trials. These conditions were designed to simulate

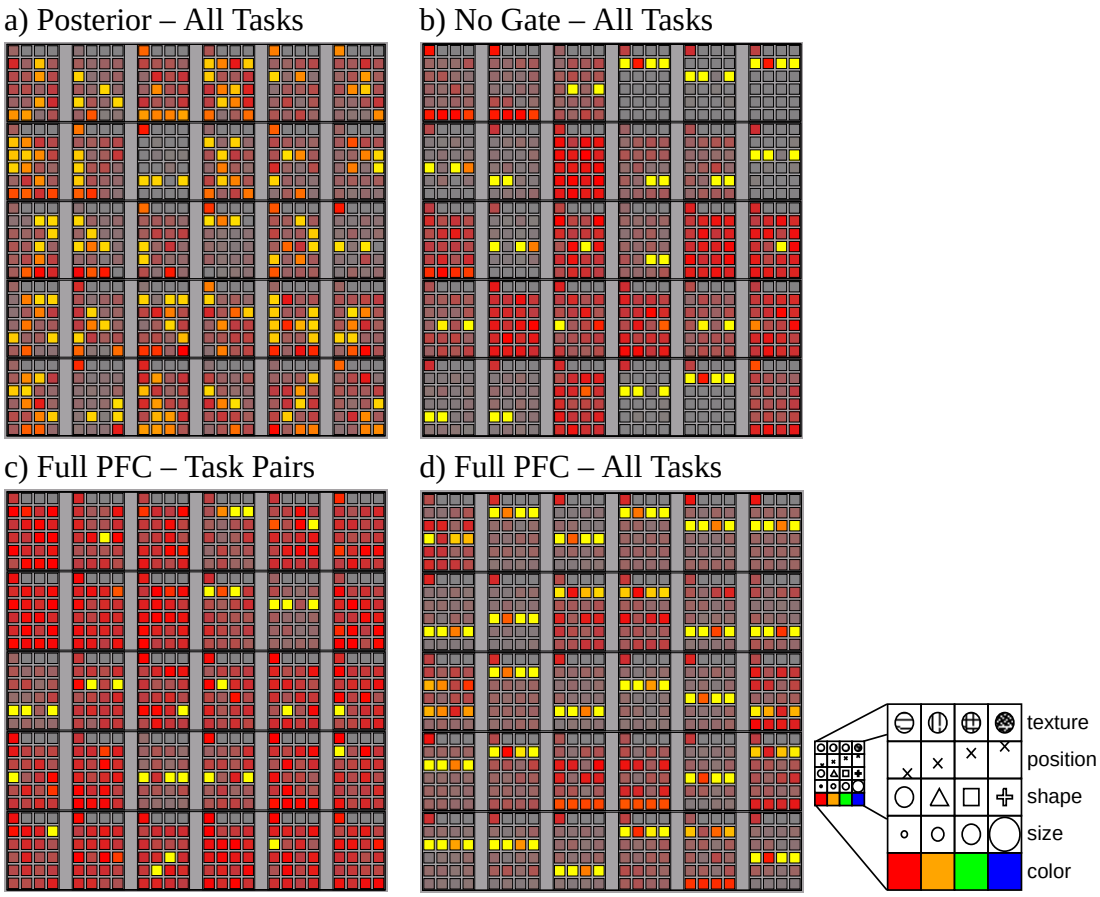

Figure 2: Representations (synaptic weights) that developed in four different network configurations: a) Posterior cortex only (no PFC) trained on all tasks; b) PFC without the adaptive gating mechanism (all tasks); c) Full PFC trained only on task pairs (NF & MF in this case); and d) Full PFC (all tasks). Each panel 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 (d) and Figure 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.

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 via a dimension cue input, to provide as generous a test as possible of models lacking the ability to maintain task-relevant information internally (see (*19,28*) for more details and effects of parametric variations).

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 (Figure 2d). That is, each PFC unit came to represent a single dimension, and to represent 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 (Figure 2b), as well as the full model trained only on task pairs (Figure 2c), and were essentially absent from the model entirely lacking a PFC (Figure 2a). 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. Note that the total number of training trials and stimulus inputs were equated across simulation conditions, so that the increased breadth of experience in the All Tasks condition was solely from exposure to more task contexts (*28*).

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 than can be achieved by dynamic updating of activation states in PFC (*27*). This difference accounts for the increased levels of perseveration observed with PFC damage in the WCST and other tasks, as has been demonstrated in several existing models (*14,15,25*), 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

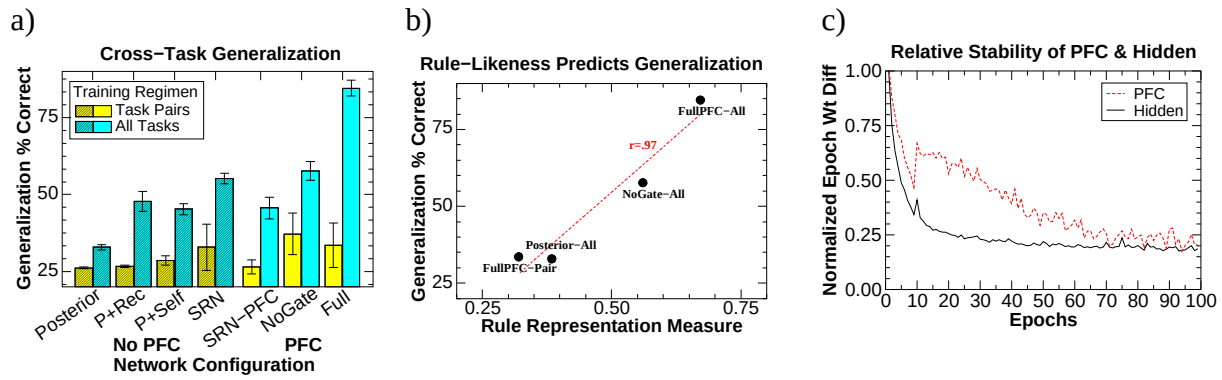

Figure 3: **a)** Cross-task generalization results (% correct on task-novel stimuli) for the full PFC network and a variety of control networks, with either only two tasks (Task Pairs) or all four tasks (All Tasks) used during training (N=10 for each network, error bars are standard errors). Overall, 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 4 dimensions, training represented only 31.6% (324) of the total possible stimulus inputs (1024); The roughly 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 —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 (the rule representation measure, described in (*29*)) for the networks shown in Figure 2. **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.

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 1/3 as many errors as other networks) on stimuli for which it had no prior same-task experience (Figure 3a). However, this was only the case 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 representa-

tions. Indeed, Figure 3b shows that, as we hypothesized, the degree to which different networks developed abstract dimensional representations (*29*) 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 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 condition 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 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 (*30*).

Our model makes the further 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), while the PFC did not stabilize until 70 epochs (Figure 3c). 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*).

Simulation Study 2

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 Wisconsin Card Sort Task (WCST), two tasks that have been used widely as benchmarks of prefrontal function (*31–34*). 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 prefrontal cortex (DLPFC) in humans (see (*19*) 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") while 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 (Figure 4a). 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, 27*). However, unlike these earlier models, PFC representations in our model developed through learning. Furthermore, Figure 4b shows that simulated lesions to the model's PFC layer (30% unit removal, post training) replicate the color naming impairments observed from DLPFC lesions in human patients (*34*), consistent with the observation that this PFC area supports abstract color dimension representations (*33*).

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, 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. Figure 4c shows that increasing amounts of PFC damage

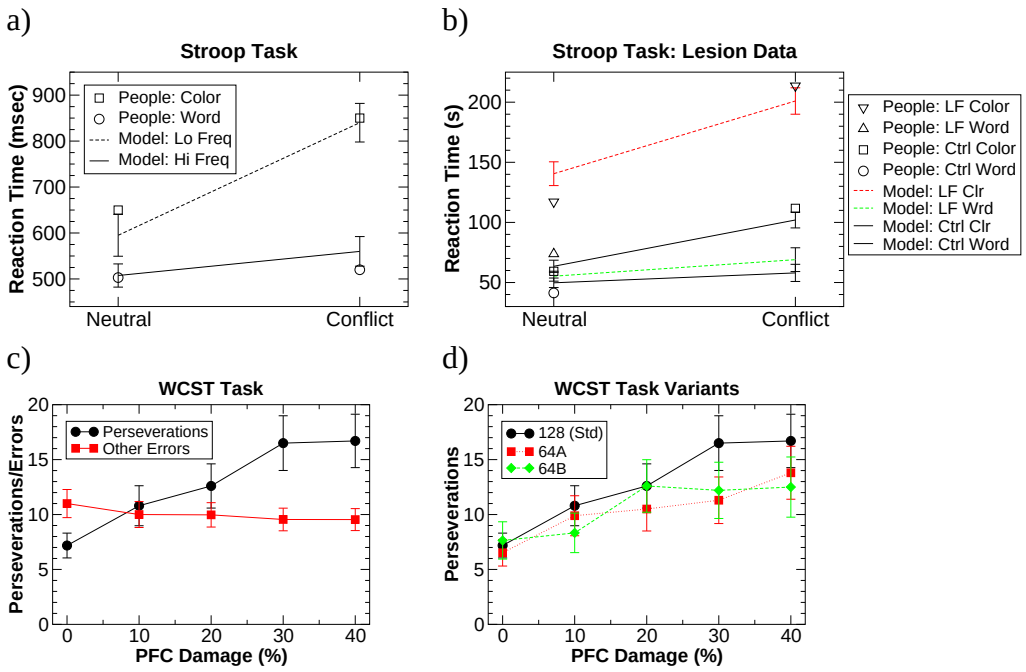

Figure 4: **a)** Performance of the full PFC network on a simulated Stroop task, demonstrating the classic pattern of conflict effects on the subordinate task of color naming with unaffected performance on the dominant word reading task (human data from (*35*)). This was simulated by training one dimension (A) with 1/4 the frequency of the others, making it weaker. In the neutral condition a single feature was active, while the conflict condition had two features present, and the dimension cue input specified which was to be named. Reaction time was measured as the number of cycles to activate a feature in the response layer $> .75$ (multiplied by 35 to match human RT's in msec). **b)** Stroop performance for a 30% lesion (removal) of PFC units in the model (post training), compared with data from (*34*) on patients with Left Frontal (DLPFC) lesions (LF) and matched controls (Ctrl) (data in seconds to complete a block of trials; model cycles were transformed with an offset of -30 and slope 5.5 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, post training). 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 (*32*) (128 is the standard case plotted before, while 64A involves providing instructions about the relevant dimensions along which cards could be sorted, and 64B has explicit instruction when the rule changes; see (*19*) for details). N=10 networks, error bars = standard error for all graphs.

(unit removal, post training) produces 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 (Figure 4d) that were observed with various levels of additional instruction provided

by Stuss and colleagues (*32*).

Discussion

The findings reported here provide new 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 cross-task 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 posterior cortex may be better at learning complex, similarity-based categories, whereas PFC can more quickly acquire simple rule-based categories (*36*). 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 long-standing "rules and regularities in language processing" debates (*37*).

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 (*27, 38*). 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 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.

References

1. P. S. Goldman-Rakic, *Handbook of Physiology — The Nervous System* **5**, 373 (1987).

2. J. M. Fuster, *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe, 3rd Edition.* (Lippincott-Raven, New York, 1997).

3. E. K. Miller, J. D. Cohen, *Annual Review of Neuroscience* **24**, 167 (2001).

4. T. Shallice, *From Neuropsychology to Mental Structure* (Cambridge University Press, New York, 1988).

5. J. Duncan, *Nature Reviews Neuroscience* **2**, 820 (2001).

6. I. M. White, S. P. Wise, *Experimental Brain Research* **126**, 315 (1999).

7. J. D. Wallis, K. C. Anderson, E. K. Miller, *Nature* **411**, 953 (2001).

8. K. Sakai, R. E. Passingham, *Nature Neuroscience* **6**, 75 (2003).

9. A. Diamond, P. S. Goldman-Rakic, *Society for Neuroscience Abstracts* **12**, 742 (1986).

10. P. R. Huttenlocher, *Neuropsychologia* **28**, 517 (1990).

11. J. B. Morton, Y. Munakata, *Developmental Science* **5**, 435 (2002).

12. A. Newell, H. A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

13. J. D. Cohen, K. Dunbar, J. L. McClelland, *Psychological Review* **97**, 332 (1990).

14. S. Dehaene, J. P. Changeux, *Cerebral Cortex* **1**, 62 (1991).

15. R. C. O'Reilly, D. Noelle, T. S. Braver, J. D. Cohen, *Cerebral Cortex* **12**, 246 (2002).

16. T. S. Braver, J. D. Cohen, *Control of Cognitive Processes: Attention and Performance XVIII*, S. Monsell, J. Driver, eds. (MIT Press, Cambridge, MA, 2000), pp. 713–737.

17. R. C. O'Reilly, T. S. Braver, J. D. Cohen, *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control.*, A. Miyake, P. Shah, eds. (Cambridge University Press, New York, 1999), pp. 375–411.

18. S. Monsell, *Unsolved mysteries of the mind: Tutorial essays in cognition*, V. Bruce, ed. (Psychology press, Hove, UK, 1996), pp. 93–148.

19. See supporting online material: ?

20. J. M. Fellous, X. J. Wang, J. E. Lisman, *Nature Neuroscience* **1**, 273 (1998).

21. D. Durstewitz, J. K. Seamans, T. J. Sejnowski, *Journal of Neurophysiology* **83**, 1733 (2000).

22. J. D. Cohen, T. S. Braver, R. C. O'Reilly, *Philosophical Transactions of the Royal Society (London) B* **351**, 1515 (1996).

23. S. Hochreiter, J. Schmidhuber, *Neural Computation* **9**, 1735 (1997).

24. M. J. Frank, B. Loughry, R. C. O'Reilly, *Cognitive, Affective, and Behavioral Neuroscience* **1**, 137 (2001).

25. N. P. Rougier, R. C. O'Reilly, *Cognitive Science* **26**, 503 (2002).

26. P. R. Montague, P. Dayan, T. J. Sejnowski, *Journal of Neuroscience* **16**, 1936 (1996).

27. R. C. O'Reilly, Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, Cambridge, MA, 2000).

28. The sequencing of tasks and relevant dimensions was organized in a hierarchically blocked fashion, with the outermost block consisting of a loop through all instructed tasks followed by all uninstructed tasks, with each task being performed for a block of 25 input/output trials. The relevant dimension was switched after every 2 of the outermost blocks (200 trials). This relatively low rate of switching allows the networks without adaptive gating mechanisms plenty of time to adapt to the relevant dimension. Other training schedules were explored with similar overall results, as described in the online supplemental materials. For a given task the model saw only a subset of the feature values along each dimension, and a relatively small fraction (about 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 roughly 10 epochs to achieve near-perfect performance on the training items, but we measured cross-task generalization performance every 5 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. To evaluate how the range of task experience influenced performance, the model was trained with two regimens: one involving all possible pairs of tasks, and the other involving all four tasks. We trained and tested different network configurations, in order to test the contribution made by constituent mechanisms to learning and performance. All network configurations had the same total number of processing units, in order 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 Figure 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 networks to generate statistics, and error bars in figures 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 over 40 different cognitive models from perception and attention to learning, memory, language, and higher-level cognition (*27*), 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 (*25*) and the online supplemental material.

29. The rule-like representation measure was computed by comparing each unit's pattern of weights to the set of 5 orthogonal, complete dimensional *target* patterns (i.e., the A dimension target pattern has a 1 for each A feature, and 0's 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: $d_i = \frac{\max_k |\mathbf{w}_i \cdot \mathbf{t}_k|}{\sum_k |\mathbf{w}_i \cdot \mathbf{t}_k|}$ where $\mathbf{t}_k$ is the dimensional target pattern k and $\mathbf{w}_i$ is the weight vector for unit i, and $|\mathbf{w}_i \cdot \mathbf{t}_k|$ represents the normalized dot product of the two vectors (i.e., the cosine). This value was then averaged across all units in the layer (PFC or Hidden, as shown in Figure 3) and then correlated with that network's generalization performance.

30. When the network produced an incorrect response, the adaptive gating mechanism temporarily inhibited PFC units that were active during that response, favoring activation of other units on the next trial. This implemented a simple form of search (random sampling with delayed replacement). The influence of this search mechanism interacted with the requirement that, for each task, the network had to attend to features in one dimension and ignore the others (the "rule" for the task). Together, these put pressure on the network to commit individual units to a single dimension. If a unit represented several dimensions, then even if one of these was the correct one, nevertheless that unit would often be inhibited because the other dimensions it represented were incorrect (and supported spurious responses). As a result of this inhibition,

connection weights supporting the activation of these units were weakened. In contrast, units that represented a single dimension were allowed to remain fully active as long as that was the correct dimension, and thus their connection weights were strengthened. As a consequence, the network developed PFC representations in which each unit was committed to a single dimension.

31. D. R. Weinberger, K. F. Berman, D. G. Daniel, *Frontal Lobe Function and Dysfunction*, H. S. Levin, H. M. Eisenberg, A. L. Benton, eds. (Oxford University Press, New York, 1991), pp. 276–285.

32. D. T. Stuss, *et al.*, *Neuropsychologia* **38**, 388 (2000).

33. A. W. MacDonald III, J. D. Cohen, C. S. Carter, *Science* **288**, 1835 (2000).

34. D. T. Stuss, D. Floden, M. P. Alexander, B. Levine, D. Katz, *Neuropsychologia* **39**, 771 (2001).

35. K. Dunbar, C. M. MacLeod, *Journal of Experimental Psychology: Human Perception and Performance* **10**, 622 (1984).

36. E. E. Smith, A. L. Patalano, J. Jonides, *Cognition* **65**, 167 (1998).

37. J. L. McClelland, K. Patterson, *Trends in Cognitive Sciences* **6**, 465 (2002).

38. Y. Munakata, *Developmental Science* **1**, 161 (1998).

39. Supported by ONR grants N00014-00-1-0246 and N00014-03-1-0428, and NIH grants MH64445. Last authorship reflects equal contribution; order was determined by a flip of coin. We thank Carlos Brody, Tim Curran, Michael Frank, Tom Hazy, Dave Jilk, Ken Norman, Yuko Munakata, Alex Petrov, and members of the CCN lab for helpful comments.