Neural Mechanisms of Cognitive Control: An Integrative Model of Stroop Task Performance and fMRI Data

Seth A. Herd, Marie T. Banich, and Randall C. O'Reilly

Abstract

■ We address the connection between conceptual knowledge and cognitive control using a neural network model. This model extends a widely held theory of cognitive control [Cohen, J. D., Dunbar, K., & McClelland, J. L. On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review, 97,* 332–361, 1990] so that it can explain new empirical findings. Leveraging other computational modeling work, we hypothesize that representations used for task control are recruited from preexisting representations for categories, such as the concept of color relevant to the Stroop task we model here. This hypothesis allows the model to account for otherwise puzzling fMRI results, such as increased activity in brain regions processing to-be-ignored information. In addition, biologically motivated changes in the model's pattern of connectivity show how global competition can arise when inhibition is strictly local, as it seems to be in the cortex. We also discuss the potential for this theory to unify models of task control with other forms of attention. ■

INTRODUCTION

Flexible cognitive control over our behavior is a key part of human intelligence. In what we call here the top-down excitatory biasing (TEB) model of cognitive control (e.g., Miller & Cohen, 2001; Cohen, Dunbar, & McClelland, 1990), the prefrontal cortex (PFC) is viewed as maintaining representations that guide control of tasks. These PFC representations provide an excitatory top-down bias to groups of neurons processing task-relevant information. Because their activity is heightened relative to neurons processing task-irrelevant information, distracting information has less effect (Corbetta, Miezin, Dobmeyer, Shulman, & Petersen, 1991). This explanation is consistent with Desimone and Duncan's (1995) biased competition model of attention—TEB theory explains task control as another form of attentional control.

This theory has the virtues of simplicity and accords with a great deal of data, but it does not address the nature and origin of these task representations in the PFC. For example, in the neural network Stroop task model of Cohen et al. (1990), it is simply assumed that the PFC has existing representations tuned for the task of naming ink colors. These color-naming task representations provide extra input to the color-naming processing areas in the posterior cortex so that they out-compete the stronger (more practiced) word-reading pathway. This extra input (top-down excitatory bias) is what supports the ability to identify the ink color of a word even when that word names a different color (e.g., ''red'' printed in green) (the Stroop task).

Previous TEB models cannot account for patterns of brain activation observed in fMRI studies of the Stroop task. Of note, more activation has been observed in brain regions responsible for processing word-related information on incongruent (''red'' printed in green) than neutral trials (''lot'' printed in green) (Banich, Milham, Jacobson, et al., 2001; Banich, Milham, Atchley, Cohen, Webb, Wszalek, Kramer, Liang, Wright, et al., 2000b). Previous TEB models predict that activity in word-related brain regions should be less than that in regions processing color information, and that this effect should happen equally in incongruent and neutral trials.

To resolve this inconsistency between theory and data, we posit that category representations are involved in cognitive control. When first called upon to identify the ink color, already existing category information about ''color'' is used to guide attentional control. This information is likely not to be specific to ink color per se, but to apply to the general category of color (although with time and practice it may be honed more specifically to ink color). Interestingly, research with monkeys indicates that the same areas of the PFC that are involved in top-down control are also involved in category representations. For instance, neurons in the PFC of monkeys distinguish between category boundaries (e.g., dog vs. cat) and the response of these neurons

University of Colorado Boulder

© 2006 Massachusetts Institute of Technology

to different category boundaries can be influenced by experience and/or practice (Freedman, Riesenhuber, Poggio, & Miller, 2002).

To investigate the possibility that attentional control may be influenced by prelearned categories or concepts, we used a model similar to that of Cohen et al. (1990). However, critical to our revised model, we added a general, abstract representation of color (which subsumes both linguistic representations of color, as manifested in words, and the perceptual representation of color, as manifested in ink colors). This inclusion changed the dynamics of activation for the task-irrelevant pathways (i.e., word reading in the incongruent condition). Critically, these different activation dynamics now match those recorded in several fMRI studies of the Stroop task, where it was observed that the activation of word-reading areas in the posterior cortex actually increased when attention should be more strongly directed away from word information (i.e., on incongruent as compared to neutral trials) (e.g., Milham, Banich, Claus, & Cohen, 2003; Banich, Milham, Jacobson, et al., 2001; Banich et al., 2000b). In contrast, the original Stroop model (Cohen et al., 1990) predicts that the top-down bias for color naming causes the activation of the competing word-reading pathway to decrease in activity. Under such a model, activation in the word-reading pathway would decrease, not increase, in the case of incongruent trials. By including abstract task representations, our model is able to reconcile seemingly inconsistent fMRI data with the widely accepted TEB model of cognitive task control.

In addition, TEB theory holds that a top-down excitatory bias onto one processing pathway in the posterior cortex will cause other areas to be inhibited. In prior neural network models, this inhibition was simulated by including direct inhibitory competition among the different processing pathways. But the reading and color-naming pathways we simulate here are in different cortical areas, whereas inhibitory projections in the human brain are strictly local. Thus, in the present model, we included only excitatory projections between the different processing pathways. These excitatory projections excite corresponding representations in other brain areas (e.g., the word ''green'' in the word-processing pathway excites the color green in the color-processing pathway), while also exciting inhibitory interneurons within each of these areas (e.g., excitation of the representation of green in the color-processing region inhibits activation of representations of other colors, such as blue and red). The net effect of this pattern of connectivity is functionally similar to the global inhibition implemented in previous models, but is more consistent with the known properties of cortical connectivity. In summary, the model presented here represents a significant advance in reconciling the TEB theory with important empirical data from neuroimaging and neuroanatomy.

Relevant fMRI Data

Many neuroimaging studies clearly support aspects of the TEB theory. For example, it has been demonstrated that there is persistent activity within the PFC during Stroop task performance in numerous studies (e.g., Zysset, Muller, Lohmannn, & von-Cramon, 2001; Banich, Milham, Atchley, Cohen, Webb, Wszalek, Kramer, Liang, Barad, et al., 2000a), as well as in other tasks requiring flexible behavior (for a review, see Smith & Jonides, 1999). fMRI evidence indicates that prefrontal activity coincides with a cue indicating task demands (e.g., a cue indicating which of two tasks should be performed). Furthermore, this activity remains at about a constant level even after stimulus presentation. In contrast, posterior regions processing task-relevant information exhibit a small increase in activation at the time of the cue (presumably as a result of top-down influences from the PFC), and a larger increase after stimulus presentation (Kastner & Ungerleider, 2000). Although those data are convincing in establishing the PFC as being the locus of maintaining task set, they do not specify how frontal activity controls performance.

Recent neuroimaging data provide new evidence on the nature of that frontal influence. In the study most relevant for the present discussion, Banich et al. (2000b) used two tasks requiring cognitive control. The first was a standard color–word Stroop task in which participants were asked to identify, via a button press, the color in which words were written. Those words named either a different color than the ink color (*incongruent* condition, e.g., the word ''red'' written in blue), or a word that had no relationship to color (*neutral* condition, e.g., the word ''life'' presented in blue). The other task was a color–object Stroop task in which participants identified (via button press) the color of a line drawing of an object. In the incongruent condition, the object was presented in a different color than that with which it is highly associated (e.g., a blue banana), whereas in the neutral condition an object associated with multiple colors (e.g., a blue car) was presented. In both of these tasks, responding was slower in the incongruent condition relative to the neutral condition.

The fMRI analysis determined those regions of increased activation on incongruent trials as compared to neutral trials. This contrast yielded activation in the dorsolateral PFC, indicating that frontal activity was significantly stronger when the task was more difficult (i.e. on incongruent trials). Of most importance, there was increased activity in a set of brain regions that have been previously identified with processing of the *to-be-ignored* dimension of the task. Within the color–word task, there was increased activity in regions of the left parietal lobe that has been associated with word processing (Jessen et al., 1999). Activity was also observed in a lateral left inferior region of the parietal lobe, as well as in a superior region of the left superior parietal

lobe (Figure 1). These areas have been identified as active preferentially when viewing and naming words versus pictures (Price, 1998), and active when words must be encoded into memory (Kelley et al., 1998). For the color–object task (again using an incongruent minus neutral comparison), there were extensive patterns of activation within the ventral visual processing stream (Figure 1), strikingly similar to those observed when objects are encoded into memory (Kelley et al., 1998).

One interpretation of this finding might be that the increased activation represents frontal inhibition of these regions. This interpretation would appear to be suggested by research emphasizing the importance of the PFC in behavioral inhibition. However, we have observed activation in these same regions for congruent trials (e.g., the word ''red'' printed in red) as compared to neutral trials (Milham, Erickson, et al., 2002), which makes this explanation unlikely. Inhibiting word information more on congruent trials than neutral trials would be counterproductive.

The model we present here shows that this counterintuitive finding can, in fact, be consistent with TEB theory, if the nature of the underlying frontal representations is more carefully considered.

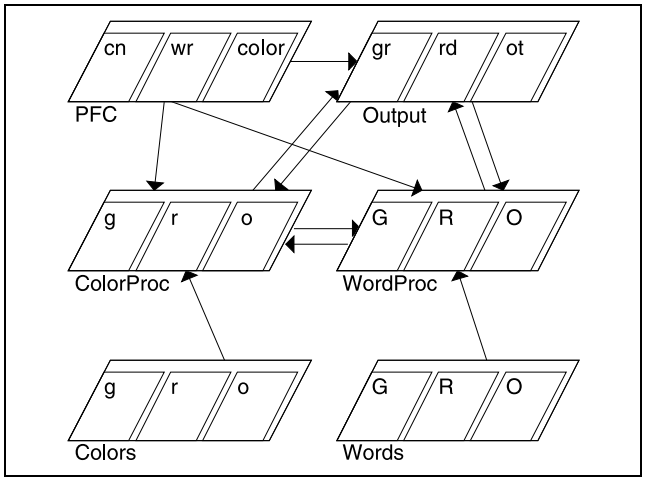

Figure 2. The model. The Colors and Words layers represent stimulus inputs, the corresponding Proc layers represent subsequent stimulus processing stages, and the Output layer represents the outcome of response selection. The PFC layer represents maintained frontal activity that appropriately biases processing to accomplish the task. Word and color streams compete or reinforce at two stages: through lateral interconnections between processing layers, and at the output layer. g = green color; r = red color; o = other color; G = word ''green''; R = word ''red''; O = other word; gr = ''green'' response; rd = ''red'' response; ot = ''other'' response; cn = color-naming task; wr = word-reading task; color = color task unit.

An Integrative Neural Network Model

The model (Figure 2) was constructed to be as similar as possible to previous models of the TEB theory (O'Reilly & Munakata, 2000; Cohen et al., 1990). To accommodate previous fMRI data and make the model biologically more plausible, two important changes were made. First, we added a third task-set unit representing the general concept of ''color'' to the two units representing the two specific color-naming tasks. Second, we added input and output units representing a noncolor word. Without these units, previous models have not provided an explanation for the fact that color words produce more interference than do noncolor words. Finally, we limited inhibitory connections to those within each region, in accord with known physiology.

The model (Figure 2) and its results can be understood entirely with ''spreading activation'' concepts: Units have an activity level and increase the activity of linked units according to that activity level and the strength of their connection. The details of the activation dynamics are discussed at the end of this section, but the model can be understood without understanding that detail.

Activation flows through the model from input layers to processing layers and finally to the single output layer.

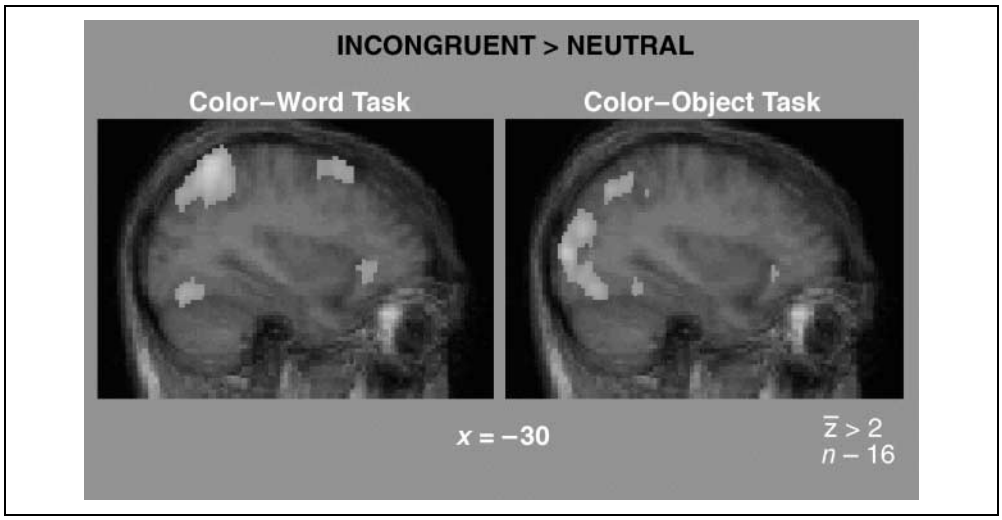

Figure 1. fMRI results from two Stroop task variants. Light areas show increased activation in the incongruent condition versus the neutral condition. In the color–word task, activation can be observed in regions that process the to-be-ignored word, a superior region of the left parietal lobe and an inferior region of the temporal lobe. In the color–object task, activation is observed in regions that process the to-be-ignored object: portions of the ventral visual processing stream.

One input layer is the ink color in which words are written, whereas the other input layer corresponds to the meaning of those words. The extra ''PFC'' layer has units that represent each task, plus the ''color'' unit that represents the general concept of color, which we hypothesize is recruited to serve as a representation for task control. These units are identical to all the others in the simulation, except that they are always held active when it is appropriate for a task. Such sustained activity has been observed in many recordings from monkey PFC and human fMRI during a variety of working memory tasks (Passingham & Sakai, 2004), and persists in the face of distracting stimuli (Miller, Erickson, & Desimone, 1996). The biological mechanism that correctly selects and maintains their activity is beyond the scope of this article; one possibility is detailed in Frank, Loughry, and O'Reilly (2001).

These ''PFC'' units effect task control through ordinary excitatory connections. Like all units, they speed and strengthen the activation of other units according to the strength of their connections. The color-naming unit is initially connected to the color-processing units (CPUs), the word-reading unit is initially connected to the word-processing units (WPUs), and the general color unit is initially connected to both. When the task is color identification, the color identification unit is automatically on for the duration of the trial. It speeds and strengthens activations of the CPUs so that they control the output units even in the presence of conflict from the WPUs.

All connection weights between units were initially set randomly, and then modified by training using Hebbian learning (O'Reilly & Munakata, 2000). The model was trained on the word-reading task five-thirds as often as it was trained on the color identification task, to approximate the amount of pre-experimental experience that a Stroop task participant has with each task. The differential training frequency is entirely responsible for establishing the dominance of word reading over color identification. On one-third of trials (two-thirds of color trials and two-fifths of word trials), congruent color and word information was presented to allow the model to develop meaningful lateral connections between word and color units. Thus, in this model, as in the original (Cohen et al., 1990), automaticity versus control is a matter of degree rather than distinguished categorically.

The model was constructed within the Leabra framework for neural network modeling (O'Reilly & Munakata, 2000; O'Reilly, 1998). This framework has been used to model a wide variety of cognitive phenomena. Because most of the parameters we used were left at the default settings (those used to model these other phenomena), the number of free parameters for this model was greatly reduced. We chose only the network configuration, gain for the unit activation function, gain for the Hebbian learning rule employed, and strengths for top-down and lateral connections. The ''out-of-the-box'' nature of this simulation indicates that the results we found are based on general principles rather than a peculiarity of our parameter choices.

The units in this framework are each modeled after single neurons. We assume that these represent an average neuron in a large population. The model uses ''point neurons'' with no spatial extent, but some biological detail. Each has a membrane potential and excitatory, inhibitory, and leak ion currents. The activation of these units is rate-coded rather than spiking, again for simplicity. The biological realism of this modeling framework is critical for some applications, but we believe that the results of the current model do not depend on the particulars of the modeling framework.

Prefrontal Representations

In short, we hypothesize that task-set representations are not constructed ''from scratch'' for the Stroop task. Instead, they are recruited from a set of existing representations that may not be perfectly aligned with the specific demands of this task. Outside of the Stroop task, people do not typically need a concept of color that excludes color words. However, with sufficient experience on the Stroop task, people may develop more appropriate task-set representations (we address this in the Discussion section).

The key to our model's explanation of the fMRI results of Banich and colleagues is that we include a generalized representation of ''color'' in the PFC task-set layer (in addition to the more specific color identification and word-reading task units used in prior models). We assume that this color representation develops in the course of normal human experience with the environment, and that it encompasses all color-related representations, both perceptual (i.e., visual colors) and linguistic (i.e., color words). Thus, when the model, and by hypothesis the participants in the Stroop task, attempt to perform color naming, they naturally activate this color representation, which has both beneficial and detrimental effects on task performance. Specifically, it facilitates performance on the congruent and neutral trials, but impairs performance on the incongruent trials, by providing top-down support to the conflicting color words.

The ''color'' task-set unit is set to the same intermediate activity level (0.5 on a scale of 0 to 1) for all trials that involve colors or color words. This follows from the hypothesis that the general concept ''color'' is a part of the task-set representation for both ''identify ink color'' and ''identify color words.'' The color identification and word-reading task-set units were active in the corresponding task conditions; these are the part of the task set specific to each task. Each was set to a level of 0.85 when active, with one exception.

The color-naming task unit was set to a slightly higher activation (1.0) during the incongruent trials. We increased its activation on these trials to correspond to

the finding of Banich and colleagues of higher frontal activation on incongruent compared to neutral trials. This extra activation was assumed to arise from a more effortful and successful implementation of task set by subjects during the more challenging incongruent condition. This practice did not produce the results we report—it worked against our principal results.

Competitive Dynamics in the Posterior Cortex

The TEB theory holds that competition takes place among the different posterior cortical representations (e.g., the word reading and color identification areas compete with each other). In the recent implementations of TEB models (O'Reilly & Munakata, 2000; Cohen et al., 1990), this competition was achieved through inhibitory connections between color- and word-processing areas. However, long-range inhibitory projections between cortical areas are rare or nonexistent (long-range cortico-cortical projections are almost exclusively excitatory; e.g., White, 1989). To remedy this problem of biological implausibility, in the present model we included only excitatory projections between the two posterior processing areas.

These excitatory connections link the competitions within each area into a global competition. This means that the excitatory connections can actually reduce activation within a layer in some cases. This happens when one representation is strong in a particular area, but a competing representation is strong in the rest of the system. Support from the remainder of the system can lead to a tied competition, so that for some time neither representation becomes fully active. In this case, the excitatory projections have had a net inhibitory effect, as inhibition is monotonically related to total input in the model, as it likely is in the cortex. This effect is crucial to task control in this model, as discussed under the heading ''Functional Distribution of Control'' in the Results section.

The response output layer in the model also includes local inhibitory competition, and this provides an important locus of response competition between the outputs of the two pathways (color identification and word reading). The top-down biasing of the color identification pathway in the incongruent trials enables it to better compete with word reading at this response output layer. However, this close competition leads to slower reaction times, as observed in the behavioral data.

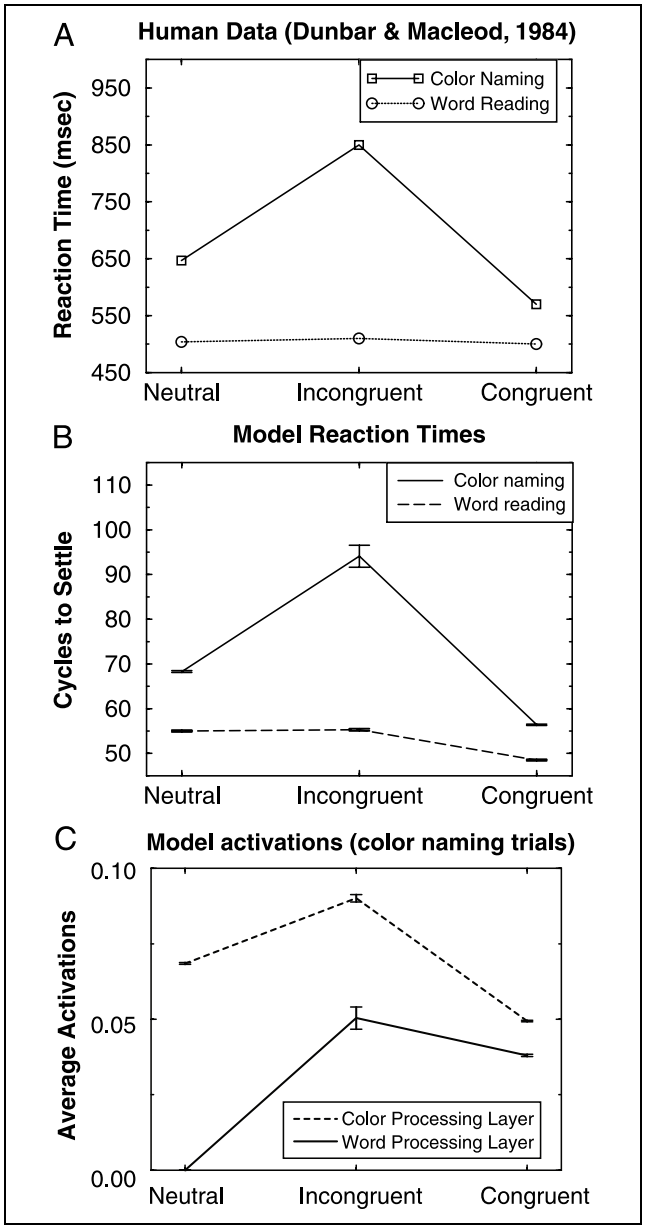

Figure 3. Human behavioral, model behavioral, and model activity level results.

RESULTS

Figure 3A and B compares the model's basic reaction time data with those from subjects—it clearly captures the basic findings as well as the previous models. However, the critical new data are shown in Figure 3C, where we plot the average activation of the units in the *word-reading* hidden layer during the three conditions of the *color identification* task. Thus, we are looking at the activation of the *irrelevant* pathway. What we see is that activation increases during the color identification incongruent condition relative to the neutral condition—this is the pattern of data reported in the fMRI studies (Banich, Milham, Jacobson, et al., 2001; Banich et al., 2000b). In general, this increased activation is attributable to the top-down excitation from the general color unit in the PFC and the between-area excitatory projections.

We can step through the flow of activation in the network to understand exactly what is happening. We use the example of an incongruent trial because the critical results are obtained there. Initially, all activation

in the processing layers and output layer is set to zero. Within the PFC layer, the color identification task-set unit and the general color task-set unit are set to activity levels of 0.85 and 0.5, respectively. This combination is the task set for the color identification task. Within the input layers, one color input unit, in this case assumed to be the ''red'' color input unit, and the conflicting ''green'' word input unit are set to a high activity level.

Activation begins to spread to the CPU representing red and the WPU representing ''green'' (r and G, respectively in Figure 2). The ''red'' CPU becomes active more quickly, because it is receiving top-down support from the color-identification PFC unit and the color PFC unit. The ''green'' WPU is also receiving top-down support due to its learned connection with the color PFC unit. This support causes it to become more active than the noncolor WPU does on neutral trials, but less active than the ''red'' CPU that receives support from both task-set units (color naming and general color). Because the red CPU becomes active both more quickly and more strongly than the green WPU, the red rather than green output unit is activated.

In summary, the reaction time differences arise from competitive interactions between the two processing pathways—the incongruent color identification condition is slower because the two pathways compete strongly in activating their associated response. Word reading is so dominant that this conflict is minimal in that condition. The color words (used during incongruent and congruent trials) receive top-down support from the general color PFC unit, whereas the ''other'' word (used during neutral trials) does not. Because the neutral word does not receive top-down support, there is less activity within the word-processing layer during the neutral condition than on the incongruent and congruent conditions. Banich, Milham, Jacobson, et al. (2001) also found word-reading area activations in comparing congruent to neutral conditions that were nearly as large as those in the incongruent to neutral comparison. This finding supports our hypothesis that semantic relation to the task set determines processing of the to-be-ignored dimension.

Effect of PFC General Color Unit

In order to determine the contribution of the general concept of color in the task set, we eliminated the general color unit at testing (eliminating the unit during training produced a failure to learn, for relatively uninteresting reasons). It might be argued that the greater activation for incongruent and congruent trials than neutral trials in the word-processing layer is driven solely by the reciprocal connections between this layer and the color-processing layer. However, the results of testing the model without the PFC general color unit showed this not to be the case. As shown in Figure 4B without the PFC color unit, activity in the word-processing layer is increased only for congruent trials but not for incongruent trials as compared to neutral trials. This effect was not driven by settling times because RT was faster on congruent trials. Without the activity of the color unit, the model failed to reproduce either the behavioral (Figure 4A) or the neuroimaging (Figure 4B) results.

It is clear that our model needs the general color task-set unit to reproduce the experimental results. However, it is possible to imagine a different model that does not. Such a model would be significantly more complex. To explain the central finding of higher word-reading activation on incongruent trials, the model would have to employ direct inhibition. It would therefore be without an easy explanation for the similar activations in the congruent condition, as inhibiting word information in that condition relative to the neutral condition would be counterproductive. The current model seems a more parsimonious explanation of the data.

Functional Distribution of Control

To test the functional mechanisms of control, we turned off learning in the projections between the separate color- and word-processing layers. This manipulation entirely eliminated task performance—the model re-

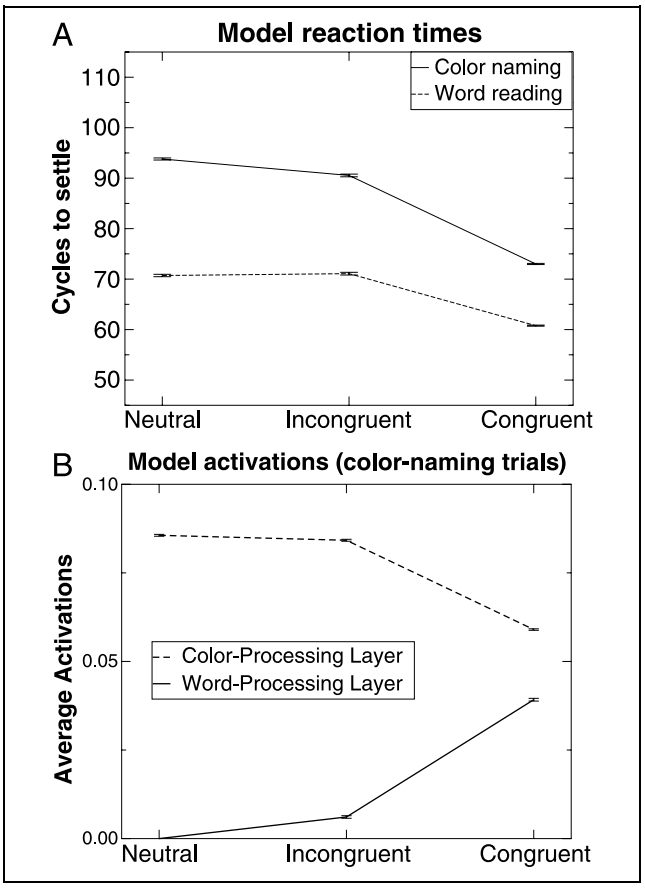

Figure 4. Model behavioral and model activity level results without ''color'' task-set unit.

sponded with the color word instead of the color in every incongruent trial, whether deactivation of the link was during training or at test only. This result shows that the excitatory projection from color processing to word processing had an inhibitory effect on that layer in the neutral and incongruent conditions.

Metaphorically, we could say that making a lesser competitor stronger can enable it to better compete, in effect inhibiting the other competitors. If auditory areas are representing a dog's bark, visual attention will focus on a dog rather than a cat. This happens without any inhibitory connection from the sound of the bark to the shape of the cat; but nonetheless, the representation for the visual form of a cat is effectively inhibited because it has less strength than the representation for the visual form of a dog.

More concretely, in this model, adding input to one unit can slow the activation of the other units, because of the competitive inhibition in each layer. The green color unit and the green word unit develop strong weights with each other because they are presented together during congruent training trials. On the incongruent trials, the green color unit becomes active more quickly (due to support from the color and color-naming units) than does the relatively unsupported red word unit. The green color unit then supports the green word unit, which competes with and slows the activation of the red word unit. The (incorrect) red word unit therefore interferes less with the output, improving control.

In addition, the red word unit's slower activation means less total activation in that layer during the trial. The excitatory projection has effectively inhibited activity in the to-be-ignored pathway. In sum, learned excitatory connections between areas encourage the system to settle into one global ''topic'' of processing, and activity in some brain regions may be temporarily suppressed while this global competition is resolved.

This paradoxical inhibitory effect may also help explain otherwise puzzling findings of apparent inhibitory relationships among cortical regions that are too widely spaced to have direct inhibitory projections between them. This result shows that excitatory projections between areas can contribute to control even when neither one of those areas contains task representations. Once top-down task control has pulled some areas to the correct topic, these areas will help pull others in line.

DISCUSSION

The model presented here accounts for the fMRI results of Banich et al., 2000b, and Banich, Milham, Jacobson, et al., 2001, that on the surface appeared to be inconsistent with the general principles behind previous neural network models of Stroop task performance (Miller & Cohen, 2001; O'Reilly & Munakata, 2000; Cohen & Huston, 1994; Cohen et al., 1990).

The explanation for the fMRI results can be summed up relatively simply: Brain areas involved in processing the task-irrelevant dimension, the word, show increased activation on incongruent trials simply because all color words have some learned connection to the concept of color. And it is this general concept of color that is invoked as a task set, as it is one that individuals have used previously.

Predictions of the Model

Perhaps the most important prediction of our model is that control representations are positive—they say what information to enhance, rather than what information to inhibit. In this view, frontal control is ''inhibitory'' in only the most general behavioral sense of the word. Calling frontal control inhibitory is potentially confusing because it implies that control represents ''what not to do,'' although there is no evidence we are aware that this is the case.

In addition, we predict that these representations are recruited from existing representations, including representations of ''concepts'' not initially constructed for use in control. Evidence of the ad hoc nature of control representations should be available, particularly early in the learning of a new task.

The model also predicts that activation of areas processing to-be-ignored items will depend primarily on whether particular distractor items are semantically related to the task being performed, rather than overall task difficulty. This prediction is supported by findings that whereas the dorsolateral PFC shows increased activation when an item captures attention due to novelty (making the task more difficult), there is no change in activation in word-related processing regions when the novel word still names a color (Milham, Banich, & Barad, 2003).

If activation is driven by semantic relation to task set, posterior word-processing areas should have nearly as high an activation in the congruent condition as they do in the incongruent condition. We expect this because the congruent color words have just as much semantic relation to the task set as do incongruent color words, but the driving control representations are slightly weaker as evidenced by less dorsolateral PFC activation in that condition. Existing data are consistent with this prediction (Milham, Erickson, et al., 2002).

A secondary prediction is that color-processing areas do have a higher activation in the incongruent than neutral or congruent conditions, despite the lack of evidence for this in the studies. However, in the model, this difference is quite modest compared to that in the word-processing layer (see Figure 3C). In general, such a difference is not observed in the contrast between incongruent and neutral trials in the imaging data. However, with training, increased activation in color-processing regions is observed (Milham, Banich, Claus,

et al., 2003), probably because the task set becomes more specific to ink color.

This empirical finding is explained by the model. It predicts that collateral activation of task-disruptive processing areas should decrease as the task representation becomes more specific. As people become more practiced, their stronger representation of the visual color can partially replace the standard representation of general color, leading to improved performance (MacLeod, 1991), and reduced activation in areas processing the to-be-ignored dimension (Milham, Banich, Claus, et al., 2003).

However, we do not predict that practice will eliminate the interference effect. Even if the task representations become perfectly precise, color words will always cause more interference than neutral words because they produce response conflict by partially activating a task-eligible response. If this prediction is true, response ineligible color words (e.g., the word ''purple'' in red ink when possible responses are red, blue, and green) should produce little interference in highly practiced participants.

Broader Implications of the TEB Model

The current model demonstrates that the recent fMRI findings of Banich and colleagues are compatible with the TEB theory of cognitive control. This theory is important beyond understanding performance of the Stroop task, as it provides a simple mechanism that could underlie all varieties of cognitive control. TEB theory provides a particularly attractive mechanistic theory of cognitive control for the following reasons.

First, control representations in TEB theory are always positive. It is, in principle, much simpler to represent a specific task, action, or stimulus to emphasize than to simultaneously represent the usually large number of competing stimuli, actions, or tasks that control seeks to de-emphasize. Second, the TEB model is computationally noncontroversial and biologically highly plausible. Although there are many conceivable mechanisms for implementing an alternative inhibitory control model, the TEB model provides a better fit with well-established biological properties of the cortex. Between-area connectivity within the cortex of higher mammals is exclusively excitatory, whereas inhibition operates locally via interneurons (e.g., Gomez-Urquijo, Reblet, Bueno-Lopez, & Gutierrez-Ibarluzea, 2000; White, 1989), as captured in the TEB model. Although current biological data do not rule out an inhibitory control model, lack of data supporting such a model is suggestive.

Another strength of the TEB theory is that it has the potential to unify attentional aspects of task control with other forms of attention. In the current model, control is the result of an appropriate bias signal influencing a global competition for representation that happens across many areas that are separate but with learned links. The highly successful biased competition model of attention in the visual system (e.g., Desimone & Duncan, 1995) could be described in exactly the same fashion, substituting ''attention'' for ''control.'' The biased competition hypothesis is expressed particularly lucidly by Duncan, Humphreys, and Ward (1997), whereas Duncan (1996) provides a wealth of converging evidence from behavioral, neuroimaging, and patient studies that supports this view. In addition, the current explanation of attentional control can be extended in a straightforward way to contextual enhancement of task performance. Representations of context could play the same functional role in guiding behavior that task-set representations play in the current model.

Conclusions

For all of these reasons, theories of the type embodied by this model seem promising. However, such theories account only for the mechanisms by which appropriate representations can control behavior. This explanation begs the question of how these mechanisms are themselves controlled: What are the neural mechanisms for the control of control? The model we present here assumes a mechanism that can activate an appropriate control representation, and maintain the activity of that representation over a relatively long period. Some initial work on this question has begun. Botvinick, Braver, Barch, Carter, and Cohen (2001) have suggested that control representations are activated in response to conflict within the cognitive system. This approach has proven fruitful in explaining experimental data, but does not address the question of how appropriate control representations are selected or develop. These questions are being addressed in models of the PFC and basal ganglia capable of learning and flexibly utilizing control representations to solve more complex tasks (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly & Frank, in press; Rougier & O'Reilly, 2002; Frank et al., 2001).

The possibility of accounting for attentional control, sensory attention, and context effects using the same neural mechanism also merits further study. We are currently developing models using this mechanism to account for attentional effects in the visual system. The possibility of a theory unifying these effects is promising and exciting. However, further investigation using neuroscience methods is necessary to empirically test the theory.

METHODS

The model was implemented using the Leabra framework, which is described in detail in O'Reilly (2001) and O'Reilly and Munakata (2000), and summarized here. Parameters were all default values except where mentioned below. These same parameters and equations

have been used to simulate over 40 different models in O'Reilly and Munakata (2000), and a number of other research models. Thus, the model can be viewed as an instantiation of a systematic modeling framework using standardized mechanisms, instead of constructing new mechanisms for each model. The model can be obtained by contacting the first author at seth.herd@colorado.edu.

Training and Connections

Each run of the model was trained until no errors were made, always less than 20 epochs. Each epoch contained 6 trials each of both types of color-reading trials (green and red), 2 trials each of both color-naming type, 4 trials each of each congruent combined type, and 10 trials of the noncolor-reading type to simulate a common word, for a total of 34 trials per epoch. The reported results are an average of 50 individual runs. The connections from the PFC layer to both processing layers and the connections between processing layers were both set to twice the relative strength of all other connections. This was empirically necessary to obtain successful task control.

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:

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

with 3 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 constant that controls the relative influence of the different conductances. This equation can also be understood in terms of a Bayesian decision making framework (O'Reilly & Munakata, 2000).

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

$$\eta_j = g_e(t) = < x_i w_{ij} > = \frac{1}{n}\sum_i x_i w_{ij} \qquad (2)$$

The inhibitory conductance is computed via the k-winners-take-all (kWTA) function described in the next section, and leak is a constant. Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

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

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $x > 0$. For this model we used $\gamma = 50$ as opposed to the default 600 to allow more graded unit response.

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 + 1th most excited unit within a layer is below its firing threshold, whereas the kth is above threshold. In the *average-based* kWTA version is the average value for the top k most excited units and is the average for the remaining $n - k$ units. This version allows for more flexibility in the actual number of units active depending on the nature of the activation distribution in the layer; we therefore used it in the processing layers with the value of the q parameter at the standard default value of 0.6.

Hebbian Learning and Weight Contrast Enhancement

For learning, Leabra uses a combination of error-driven and Hebbian learning. However, in this simulation, only the more widely accepted Hebbian learning was used. 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 (Oja, 1982).

The equation for the Hebbian weight change is:

$$\Delta_{hebb} w_{ij} = x_i y_j - y_j w_{ij} = y_j(x_i - w_{ij}) \qquad (4)$$

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

$$\Delta_{sberr} w_{ij} = [\Delta_{err}]_+(1 - w_{ij}) + [\Delta_{err}]_- w_{ij} \qquad (5)$$

Leabra implements contrast enhancement by passing the linear weight values computed by the learning rule through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left(\theta \frac{w_{ij}}{1 - w_{ij}}\right)^{-\gamma}} \quad (6)$$

Where $\hat{w}_{ij}$ is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset θ and a gain γ. Here we did not use the standard defaults of 1.25 and 6, respectively, but rather treated these as free parameters in lieu of directly adjusting frequencies of events; the final parameters used were offset $\theta = 0.75$ and a gain $\gamma = 3$.

Acknowledgments

Reprint requests should be sent to Randall C. O'Reilly, Department of Psychology, University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345, or via e-mail: oreilly@psych.colorado.edu.

REFERENCES

Banich, M., Milham, M., Jacobson, B., Webb, A., Wszalek, T., Cohen, N., & Kramer, A. (2001). Attentional selection and the processing of task-irrelevant information: Insights from fMRI examinations of the Stroop task. In C. M. Casanova & M. Ptito (Eds.), *Progress in brain research: Vol. 134. Vision: From neurons to cognition.* Amsterdam: Elsevier Science.

Banich, M. T., Milham, M. P., Atchley, R., Cohen, N. J., Webb, A., Wszalek, T., Kramer, A. F., Liang, Z. P., Barad, V., Gullett, D., Shah, C., & Brown, C. (2000a). Prefrontal regions play a predominant role in imposing an attentional ''set'': Evidence from fMRI. *Cognitive Brain Research, 10,* 1–9.

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

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

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

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

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., & Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: Functional anatomy by positron emission tomography. *Journal of Neuroscience, 11,* 2383–2402.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193.

Duncan, J. (1996). Cooperating brain systems in selective perception and action. In T. Inui & J. L. McClelland (Eds.), *Attention and performance* (pp. 85–105). Cambridge: MIT Press.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology, 7,* 255.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between the frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, and Behavioral Neuroscience, 1,* 137–160.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2002). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology, 88,* 929–941.

Gomez-Urquijo, S. M., Reblet, C., Bueno-Lopez, J., & Gutierrez-Ibarluzea, I. (2000). Gabaergic neurons in the rabbit visual cortex: Percentage, layer distribution and cortical projections. *Brain Research, 862,* 171–179.

Jessen, F., Erb, M., Klose, U., Lotze, M., Grodd, W., & Heun, R. (1999). Activation of human language processing regions after the presentation of random letter strings demonstrated within event-related functional magnetic resonance imaging. *Neuroscience Letters, 270,* 13–16.

Kastner, S., & Ungerleider, L. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23,* 315–341.

Kelley, W. M., Miezen, F. M., McDermott, K. B., Buckner, R. L., Raichle, M. E., Cohen, N. J., Ollinger, J. M., Akbudak, E., Conturo, T. E., Synder, A. Z., & Petersen, S. E. (1998). Hemispheric specialization in human dorsal cortex and medial temporal lobe for verbal and nonverbal memory encoding. *Neuron, 20,* 927–936.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203.

Milham, M., Banich, M., & Barad, V. (2003). Competition for priority in processing increases prefrontal cortex involvement in top-down control: An event-related fMRI study of the Stroop task. *Cognitive Brain Research, 17,* 212–222.

Milham, M., Banich, M., Claus, E., & Cohen, N. (2003b). Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. *Neuroimage, 18,* 483–493.

Milham, M. P., Erickson, K. I., Banich, M. T., Kramer, A., Webb, A., Wszalek, T., & Cohen, N. J. (2002). Attentional control in the aging brain: Insights from an fMRI study of the Stroop task. *Brain Cognition, 49,* 277–296.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.

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

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15,* 267–273.

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

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

O'Reilly, R. C., & Frank, M. J. (in press). Making working memory work: A computational model of learning in the frontal cortex and basal ganglia. *Neural Computation.*

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

Passingham, D., & Sakai, K. (2004). The prefrontal cortex and working memory: Physiology and brain imaging. *Current Opinion in Neurobiology, 14,* 163–168.

Price, C. J. (1998). The functional anatomy of word comprehension and production. *Trends in Cognitive Science, 2,* 281–288.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences, U.S.A., 102,* 7338–7343.

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

Smith, E. E., & Jonides, J. (1999). Neuroscience: Storage and executive processes in the frontal lobes. *Science, 283,* 1657.

White, E. L. (1989). *Cortical circuits: Synaptic organization of the cerebral cortex, structure, function, and theory.* Boston: Birkhäuser.

Zysset, S., Muller, K., Lohmannn, G., & von-Cramon, D. Y. (2001). Color–word matching Stroop task: Separating interference and response conflict. *Neuroimage, 13,* 29–36.