Accounting for Stroop Task Neuroimaging Data: Lateral Interactions & Frontal Representations

Seth A. Herd
Department of Psychology
University of Colorado Boulder
345 UCB
Boulder, CO 80309
sethherd@psych.colorado.edu

Randall C. O'Reilly
Department of Psychology
University of Colorado Boulder
345 UCB
Boulder, CO 80309
oreilly@psych.colorado.edu

Abstract

This work integrates new neuroimaging results with existing behavioral and computational studies to produce a model of attentional control within the Stroop task. The Stroop task is the paradigmatic example of how attentional control can facilitate processing of weak, less well-practiced behaviors, such as the ability to name the ink color of a word in the face of the more dominant response of reading the word itself. Attentional control is implemented in the model as an excitatory bias from a maintained frontal task set interacting with competitive dynamics between posterior cortical pathways that process relatively domain-specific information (e.g., ink color vs. words). Our model advances previous models by including lateral excitatory and inhibitory connectivity (via interneurons) based on biological principles. We show that the complex dynamics supported by these lateral connections, together with assumptions about the nature of frontal task representations, can produce counter-intuitive patterns of neural activation that have been observed in fMRI studies of the Stroop task, which appeared to contradict existing models. In sum, we show that the neuroimaging data support a model of attentional control as a top-down excitatory bias coupled with competition for activation within and between cortical areas.

1 Introduction

People are capable of performing a multitude of tasks, ranging from stepping on the brakes in response to a red light, to composing a symphony. Performance of these tasks is driven by complex interactions of external cues and internal goals. The Stroop task is the paradigmatic laboratory task for studying how behavior can be flexibly controlled by an interplay of external and internal factors. In this task, subjects need to overcome a dominant stimulus-driven response of reading a word in favor of responding to the ink color that the word is printed in — this ink color response must be supported by internal goals because the stimuli are ambiguous. We build upon the basic ideas developed in the original Stroop model of Cohen, Dunbar, and McClelland (1990) as implemented in O’Reilly and Munakata (2000). Specifically, the original model has internal goals encoded via task units,
thought to represent the activity of prefrontal cortex neurons, that provide top-down support for the ink color processing pathways in the posterior cortex. This top-down support is needed to counter stronger bottom-up (stimulus driven) processing of word information, which is built up over our extensive experiences with reading.

The implementation by O’Reilly and Munakata (2000) leveraged the biologically-based Leabra neural network modeling framework to include bidirectional and competitive activation dynamics (see also Cohen & Huston, 1994). These models, despite their simplicity (12 units each) are able to explain most of the 18 empirical results identified by MacLeod (1991) in his comprehensive summary of Stroop research. However, both of these models are unable to account for recent fMRI results obtained by Banich and colleagues (Banich, Milham, Jacobson, Webb, Wszalek, Cohen, & Kramer, in press; Banich, Milham, Atchley, NJ, Webb, Wszalek, Kramer, Liang, Wright, Shenker, & Magin, 2000b). According to the logic behind these models, a frontal task set supports the activity of the neural populations performing that particular task, allowing them to become more active than competing populations performing other tasks, and thereby win in a competition for response outputs. If this logic holds true, we would expect to see more activation of color processing areas and less activation of word reading areas when frontal support for the color naming task is strongest. This prediction seems to be directly violated in the fMRI results.

1.1 Relevant fMRI Data

First of all, it is clear that much neuroimaging data supports at least part of the theoretical framework argued for by Cohen et al. (1990). It has been convincingly demonstrated that there is persistent activity within the prefrontal cortex (PFC) during Stroop task performance (Banich, Milham, Atchley, Cohen, Webb, Wszalek, Kramer, Liang, Barad, Gullett, Shah, & Brown, 2000a; Zysset, Muller, Lohmann, & von Cramon, 2001), as well as in other tasks requiring flexible behavior (for a review see Smith & Jonides, 1999). While those data are convincing in establishing the frontal locus of a maintained task set, they do not specify how frontal activity controls performance. Banich and colleagues (Banich et al., in press; Banich et al., 2000b) provide valuable evidence on the nature of frontal influence. They present a somewhat paradoxical finding of greater activation of brain areas associated with processing of the irrelevant dimension on trials where that irrelevant dimension contained color information that conflicted with the color to be named. They used two tasks. The first was a standard Stroop task in which participants were asked to identify the color in which words were written, both when those words named a different color (conflict condition, e.g., the word red in blue ink), and when the words named a color neutral object (neutral condition, e.g., the word life in blue ink). The other task presented participants with a colored line drawing of an object either associated with a different color than that it was presented in (conflict condition, e.g., a blue banana) or associated with multiple colors (neutral condition, e.g., a blue car).

In both color-word and color-object tasks they used a comparison of conflict trials versus neutral trials. They observed increased dorsolateral PFC and anterior cingulate cortex activation in conflict trials, indicating that frontal task set activity was significantly stronger in the more difficult conflict condition than in the neutral condition. However, they showed no increase in activity in color processing regions such as area V4 in the conflict condition. Furthermore, they did find increased conflict activity in a set of areas that have been previously identified with processing of the to-be-ignored dimension of the task (Figure 1). Within the color-word task, they observed increased activity in the left precuneus region of parietal lobe, which has been shown to activate preferentially when processing words versus letter strings. Activity was also observed in a lateral left inferior region of the parietal lobe, as well as a superior region of left superior parietal lobe, which have been identified as active preferentially when viewing and naming words versus pictures and active when words must be encoded into memory. Within the color-object task, (still using an conflict
minus neutral comparison), Banich et al. observed “extensive patterns of activation noted within ventral visual processing stream ... strikingly similar to those areas activated when objects are encoded into memory.”

The data obtained by Banich and colleagues therefore seems to indicate that there is greater activation of brain areas processing the to-be-ignored dimension when frontal control is strongest, apparently contradicting the Stroop models of attentional control. These findings seem to support an interpretation of frontal control in which that control directly inhibits the to-be-ignored areas, perhaps producing the observed increased blood oxygen level due to inhibitory activity co-existing with excitatory activity in those areas as suggested by Banich and colleagues. We demonstrate instead that this activity can be explained within the overall framework of the Stroop models in terms of collateral activation of the to-be-ignored areas when they process a stimulus with associations to the maintained task set (i.e., color).

2 An Integrative Neural Network Model

We present here a model that accounts for both the behavioral results obtained in the Stroop task, and the fMRI results just described. The current model is based in part on the earlier models by Cohen et al. (1990) and O’Reilly and Munakata (2000). Like the O’Reilly and Munakata (2000) model, the present model was constructed within the Leabra framework for neural network modeling, which includes bidirectional and competitive activation dynamics, together with both error-driven and Hebbian learning (O’Reilly, 1998; O’Reilly & Munakata, 2000). However, it differs from this model in two key respects. First, it uses a more biologically-detailed mechanism for competitive inhibition, based on known properties of cortical long-range connectivity. Second, it incorporates more complex assumptions regarding the nature of the prefrontal representations that are activated to bias processing in posterior cortical areas. Before describing these differences in detail, we review the basic architecture of the model.

The model is shown in Figure 2. It has color and word sensory inputs (bottom layers) that project into corresponding posterior cortex layers. The neutral condition is simulated by activating a “neutral” (non-color) word written in a particular ink color, and other conditions are simulated by combinations of color ink and word inputs. The prefrontal cortex (PFC) has units that project excitation to each of the posterior layers (i.e., top-down biasing), and a color unit that projects to units in both layers that process color information. The output
Figure 2: The model. Input layers represent stimuli; processing layers represent domain-specific posterior cortical areas; output layer represents the initiation of a motor response, and PFC layer represents maintained frontal activity that appropriately biases processing to accomplish the task. Word and color processing compete and reinforce through lateral interconnection and at the output.

contains responses the network can make. The model was trained on the word reading task three times more often than it was trained on the color naming task, to approximate the amount of pre-experimental experience that a Stroop task participant would have with each task. This is entirely responsible for establishing the dominance of word reading over color naming, showing as in the original Cohen et al. (1990) model that automaticity vs. control is a matter of degree based on experience.

2.1 Competitive Inhibition and Long Range Projections

In the basic Leabra algorithm, and the O’Reilly and Munakata (2000) Stroop model, competitive inhibition is implemented via a k-winners-take-all approximation (kWTA), instead of explicitly simulating the inhibitory interneurons that actually perform this function in the brain. This kWTA approximation offers a number of important performance advantages, at the cost of not capturing some of the more complex and subtle dynamics that are likely to hold in the brain. The present model provides an example of such dynamics, by replacing the kWTA function with explicitly simulated inhibitory interneurons and associated long-range excitatory cortical projections. It is a well established fact of cortical anatomy that most long-range cortico-cortical connectivity is carried by excitatory pyramidal axons originating and terminating in the superficial layers (2-3) (e.g., White, 1989). Furthermore, these long-range excitatory connections synapse on both pyramidal excitatory and inhibitory local circuit neurons. Thus, long-range connections exert a mixture of both specific excitatory and general inhibitory influence.

The model has two posterior cortex layers, one for color processing and the other for word reading. Each of these layers projects both to the other layer, and to the inhibitory unit that controls that layer’s activity. Therefore, any activity within one layer has a tendency to indirectly suppress both other activity within that layer, and activity within the other layer. This inhibitory dynamic is critical to the overall performance of the model. It is important to note, however, that the interaction between layers is not of a purely inhibitory nature. The projections also have the potential to encourage the activity of a correlated representation in the other hidden layer. For example, the green unit in the color processing layer tends to encourage the activity of the green unit in the word processing layer. Clearly, this kind of associational activation makes clear functional sense (e.g., according to standard spreading activation models), and it also converges with the biological features just reviewed. We also
believe that such interaction is critical in giving rise to many attentional effects through the dynamic evident in the model of both inhibitory and excitatory net effects depending on the amount of correspondence between the topics of processing within the two areas.

2.2 Prefrontal Representations

The previous models assumed that the frontal cortex maintains a very specific and appropriate task set, activating only the posterior areas relevant to the current task. Instead, we have assumed that frontal representations are largely recruited from pre-existing learned representations of concepts. According to this view, a variety of representations associated with color, including those not specific to naming an ink color, will be activated in the color-naming condition. The model therefore has three frontal units, one that represents the frontal activity that uniquely supports the color naming task, one that represents frontal activity that uniquely supports the word reading task, and one that represents the general concept of color, whose effect is to support the activity of both word-reading and color-naming units associated (through training) with the general concept of color. The activity of the PFC units was externally set to a theoretically appropriate level on each trial. The mechanism underlying persistent activation of task representations in frontal cortex is beyond the scope of the current model, and was simply assumed. For some insight into possible mechanisms, see Frank, Loughry, and O’Reilly (in press).

The frontal units were given a slightly higher activation during the conflict trials to correspond to the finding of Banich and colleagues of higher frontal activation on conflict compared to neutral trials. This extra activation produced an effect counter to finding interference in the conflict color reading trials, and was implemented strictly to remain in close accord with relevant fMRI results. This extra activation was assumed to arise from greater effort by subjects during the more challenging conflict condition.

3 Results

Figures 3a and b compare 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 is 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 naming task. Thus, we are looking at the activation of the irrelevant pathway. What we see is that activation increases during the color naming conflict condition relative to the neutral condition — this is the pattern of data reported in the fMRI studies (Banich et al., 2000b; Banich et al., in press). In general, this increased activation is attributable to the “long-range” excitatory interactions between the color naming and word reading pathways and to the general top-down activation from the general color unit in the PFC. More specifically, we can step through the flow of activation in the network to understand exactly what is happening.
We use the example of a conflict 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 naming task set unit, and the general color task set unit are activated. 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 active. Activation begins to spread to the red color processing unit and the green word processing unit, and the red color processing unit becomes active more quickly, because it is also receiving top-down support from the color-naming PFC unit and the color PFC unit. However, this red color processing unit also sends excitation to its associated red word processing unit, contributing overall activation to the other layer. Furthermore, both the red and green word processing units are also receiving top-down support due to their learned connections with the color PFC unit. Therefore, the overall activation of the word processing layer is greater in this conflict condition than it is in the neutral condition where a non-color word is activated, which does not receive either form of extra activation.

The red color naming unit nevertheless has the greatest overall activation as it is more directly supported by the PFC, and when it becomes strongly active it inhibits activity of other units in both processing layers by activating both inhibitory units. Because the red color processing unit becomes active both more quickly and more strongly than the green word processing unit, it drives the response of the output layer to activate the red output unit. Because the green word processing unit is also active to some degree, it will partially activate the green output unit, which competes with the red output unit, making it activate more slowly than it would if no competing input were present.

In summary, the reaction time differences arise from competitive interactions between the two processing pathways — the conflict color naming condition is slower because the two pathways compete strongly in activating their associated response. Word reading is so dominant that this conflict is minimal for the word reading task. The fact that the color word receives both top-down support from the general color PFC unit and collateral activation from the color processing pathway gives rise to the finding of greater activity within the word processing layer during the conflict condition than during the neutral condition, which simulates the fMRI results.

4 Discussion

The most important contribution of the present model is that it accounts for patterns of neural activation in fMRI studies (Banich et al., 2000b; Banich et al., in press) that on the surface appeared to be inconsistent with the general principles behind the previous Stroop models (Cohen et al., 1990; O’Reilly & Munakata, 2000). Specifically, these fMRI studies appeared to suggest that the PFC is contributing specific inhibition of to-be-ignored processing areas, and this inhibition produced greater levels of BOLD signal. Instead, we showed that this activation pattern is consistent with PFC top-down excitatory biasing of task-relevant processing areas, which have a complex balance of excitatory and inhibitory connectivity with the task-irrelevant processing areas. Furthermore, the PFC representations may not be so perfectly tuned to these arbitrary laboratory tasks (having been shaped by real-world experience), such that they can provide some degree of more general activation even to task-irrelevant areas.

More generally, the principles behind our model are in accord with a broadly-supported theory of attentional control as a top-down excitatory bias to appropriate task processing populations (Miller & Cohen, 2001; O’Reilly, Braver, & Cohen, 1999; O’Reilly & Munakata, 2000). This view is attractive because it is easily integrated with a view of lower-level attentional effects as arising from appropriate learned top-down and contextual biasing within an environment of competition for representation (Desimone & Duncan, 1995; O’Reilly & Munakata, 2000). The most interesting aspect of this explanation is its
very simplicity. The phenomena of attentional control are represented simply as top-down excitatory connections, biasing one task to dominate others in a competition to influence responding. Within this explanation, support for a particular task can also arise from incidental collateral connection, providing a ready explanation for the wealth of context effects seen in cognitive tasks, as well as an explanation for non-effortful yet situationally appropriate behavior. Unlike some other theories of attention, this view requires no complex specialized attentional or control mechanism beyond a mechanism to maintain appropriate frontal activity, theorized to be adapted from a mechanism for physical action (Frank et al., in press).

To advance this theory further using neural network models, it will be essential to develop more complex, larger-scale models that address a wider range of complex attentional phenomena (e.g., visual search). We are currently working on extensions of the present model based on large numbers of units within each layer, employing distributed representations, and plan to address visual search phenomena in subsequent models.

5 References


