**A Computational Model of Cognitive Processing during Response Conflict:**

**Evidence from a Three-layer Neural Network**

**Carlos R. Salas**

**General Overview:**

The current implementation of a three layer neural network aimed to simulate the following aspects of the human cognitive system: a) the effect of differential practice, b) attention allocation, c) response conflict resolution, and d) differences in responses as a function of stimulus set. The model was able to generate evidence of each of these and demonstrates how they change across the processing time-course.

**Brief Theoretical Background:**

Conflict in the cognitive and motor system can occur whenever an external or internal cue elicits more than one possible response. These cues can be imposed by the task environment as constraints (e.g., task instructions requesting a particular type of response), or by the internal structure and the given state of a system. Practice and exposure effects can play a particularly important role in shaping the system’s structure, since they can lead to the development of prepotent responses if the system is adaptive. Prepotent responses can be the result of routine practice (e.g., reading, etc.) or can be instinctively built into the system (e.g., recoiling from the sudden heat of a fire, etc.). In the present simulation, we will focus on the former by simulating performance on the Stroop task.

The Stroop task has long been used as a classic measure of response conflict, which it induces by manipulating aspects of both the task instructions and the stimuli presented. In the Stroop task, participants are either asked to respond with the color of the ink that a word is written in, or respond by reading the word itself (e.g., and ignoring the ink color). This helps shift attention to certain features of the stimuli in the task environment. Participants tend to respond more *accurately* and *faster* whenever the color of the ink and the word match (e.g., the word RED written in red ink), relative to when the ink color and word do not match (e.g., the word RED written in green ink). This is consistent with the current conception of response conflict. Furthermore, participants tend to be faster when the task instruction is to read the word relative to when they’re asked to identify the ink color. Given that participants have had much more practice reading text than naming colors, this speed-up is consistent with the idea of reading being a more prepotent response.

**Model Implementation:**

Stroop task performance was simulated by constructing and training a three-layer neural network (see the network architecture diagram that follows). This neural architecture was chosen based on previous models in the PDP literature (Botvinick et al., 2001; Kanne, et al., 1998) and was structured to be the simplest model necessary to capture the Stroop task environment and instructional demands. Information from the task environment and task instructions was processed by propagating activation across the model’s three layers (i.e., from Input, to Hidden, to Output, and eventually to the Response Mechanism).

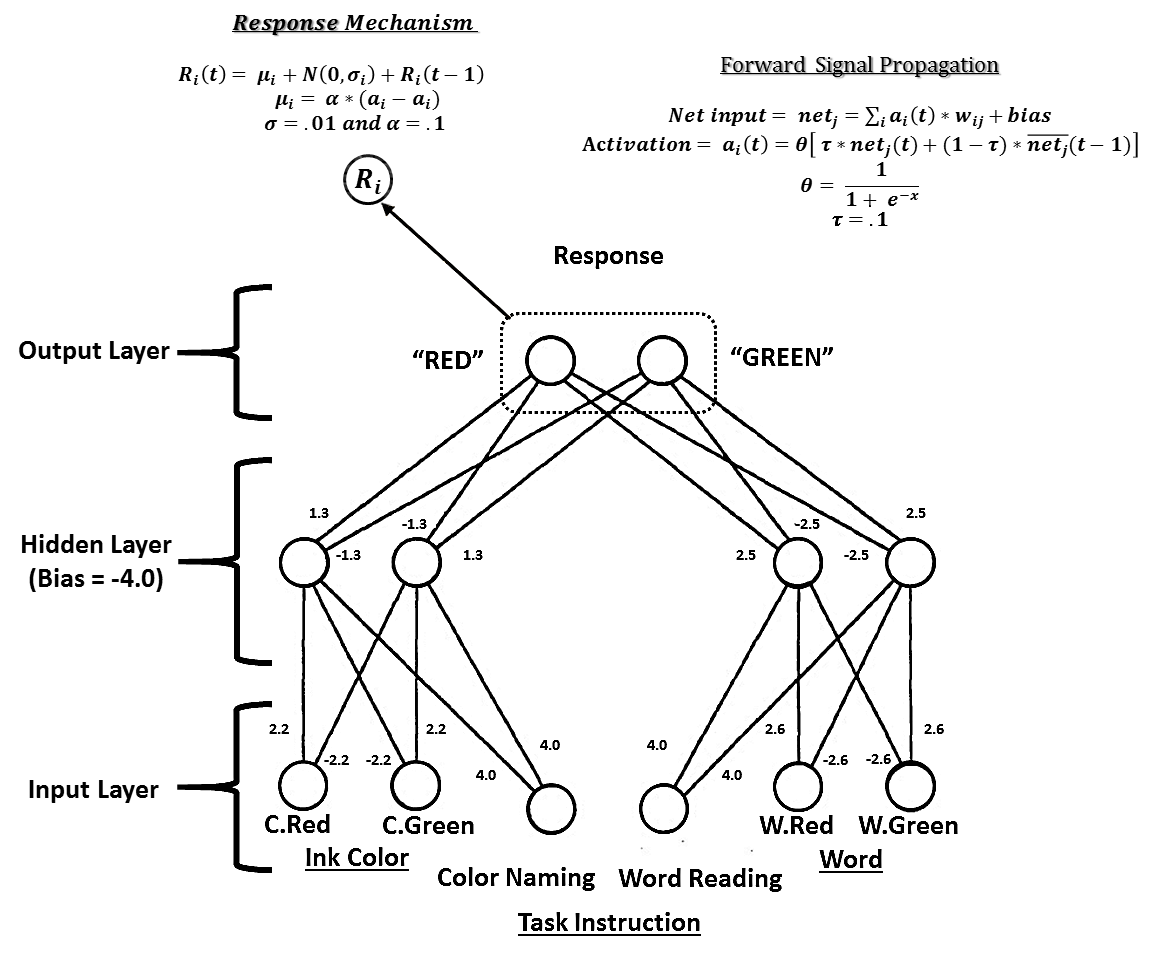
The Input layer nodes received a vector of binary values (0, 1) representing the following potential stimuli features: ink color (Red vs Green), task instruction (Color Naming vs Word Reading), and the Word (Red vs. Green). Activation from these then flowed to the hidden layer. The Hidden layer represents the intermediate nodes between the Input and Output layers. There are four nodes in this layer: two nodes in the Color pathway, and two in the Word pathway. Each node is linked and thus receives signals from three of the nodes from the Input layer. For the nodes in the Color pathway, the links fed in from both the Red and Green Color nodes, plus the Color Naming task node. For the nodes in the Word pathway, the links fed in from both the Red and Green Word nodes, plus the Word Reading task node. Finally, the Output layer had two nodes, one for each color response (i.e., Red and Green). These nodes received input from every node in the Hidden layer. The co-activation of each of the Output nodes per cycle was fed into a response mechanism (specified below) to calculate when a response was given by the system.

Information processing in the network was run in cycles. Each cycle involved spreading activation from the Input nodes through the Hidden nodes, and into the Output nodes. The equation for forward signal propagation is detailed in the network diagram, but in short summary, signals (i.e., “net inputs”) were propagated by summing the product of the activations and weights for all nodes connected to the focal node in the current layer, plus bias (if any). Activation for a given node was then calculated by taking the node’s net input, adding it to that node’s net average from the previous cycle, and then weighting it by a cascade rate (set to .1 for all layers). Gaussian noise was added to this sum, which was then fed through a sigmoidal function. At the end of a cycle, activation from the two output nodes was tallied by an evidence accumulator node in the response mechanism. This evidence accumulator node calculated a running total from the previous and current activation values from each output node. Evidence for a particular node was calculated by adding the product of an evidence accumulation rate and the difference between the activation of that focal node *ai* and its competitor *aj*. The equation for this response mechanism is detailed in the network diagram below. A processing trial ended and the response was recorded whenever the evidence accumulator for a given node reached a threshold activation of 1.0.

The network was initially trained to read a set of color words, and to identify the color of a word's printed ink. Different amounts of exposure to words and color patterns were given in order to mimic human's overlearned tendency for word reading relative to color naming (ratio was 5:1, respectively; see Kanne et al., 1998). The training resulted in updating of the network's node connection weights, which was accomplished via the backpropagation algorithm using gradient descent. The network’s weights were initially set to random strength values, but after training, the resulting weight matrix suggested a network with two processing sub-system pathways: a Color Naming pathway, and a Word Reading pathway (see Cohen, Dunbar, & McClelland, 1990).

Simulating performance on the Stroop task occurred in two phases. Phase one involved priming the network. I used the approach suggested by Mewhort et al., 1992 (initially generated by Cohen, Dunbar, & McClelland, 1990) and set all the values in the active task path and outputs to .5, and all others to .01. This task priming phase simulates the instructions given to participants. Phase two involved actually running the trial simulation by feeding in a vector of inputs, letting activation propagate through the network across cycles, and recording the number of cycles taken to generate a response. Cycles were converted to millisecond reaction time equivalents via the following equation as suggested by Cohen & Huston (1994): RT = .98\*(cycles) + 298.

**Network Architecture**



**Flow of Activation**

**Input for the Trial Simulations reported in the Run Documentation**

The following represent the input vectors fed into the model for the simulation in the sample runs document. The consistent condition simulates the model being asked to read the word RED printed in Red ink. The inconsistent condition simulates the model being asked to read the word RED printed in Green ink.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Stimulus Condition** | **Ink Color** | | **Task Instruction** | | **Word** | |
| Red | Green | Color Naming | Word Reading | RED | GREEN |
| Consistent | 1 | 0 | 0 | 1 | 1 | 0 |
| Inconsistent | 0 | 1 | 0 | 1 | 1 | 0 |

**Results**

The current model sought to capture the effect of differential practice. As is the case with human learning, the model adapted to the greater amount of exposure to certain stimuli. Structural evidence of this can be seen in the cross-layer connection weights in each relative pathway. Specifically, the connections between the neurons in the Word reading pathway are stronger than in the Color naming pathway. This was achieved with the backpropagation algorithm and differential exposure to the stimuli (i.e., via the training sets) in manner similar manner to what humans participants receive throughout their life and education (e.g., a higher proportion of time spent reading words rather than naming colors). This effect is consistent with the role of practice effects and learning.

The Run Documentation file contains several additional pieces of evidence suggesting that the model is adequately simulating human performance. The first page includes the activation for the neurons in the hidden layer. The top row of plots all show the activation for the neurons in the Color pathway, which simulates color recognition. As can be seen, in the consistent condition (two plots on top left), hidden color neuron #1 (HC1) is more active than hidden color neuron #2 (HC2), which reflects the weights that it has with the color input nodes (see diagram above), and suggests that the red ink is passing activation along the appropriate pathway. It also foreshadows (i.e., given the excitatory and inhibitory path weights from each neuron to the output layer) that the RED output neuron will receive plenty of activation from the color pathway in the hidden layer, and the GREEN output neuron will receive less. In the *inconsistent* condition (top two plots on the right), these relationships are reversed, with HC1 showing less activation than HC2. This again shows that the two hidden nodes in the pathway are being appropriately activated by the green color ink in the input, and foreshadows that this layer will pass less information to the RED output neuron than to the GREEN pathway.

The bottom row of plots displays the activation for neurons in the Word pathway, which simulates word reading. Given that the instruction and the word was the same for both consistent and inconsistent conditions, the pattern of activation for both of the neurons in the hidden layer is very similar (i.e., HW1 > HW2), despite some stochastic variation across the neurons’ firing rates. Note also that the HW1 has the most unique signature, with a fast spike that asymptotes higher than any of the other neurons in the hidden layer. This reflects its afferent signals from the Word Reading and the red Word nodes. Also note that the activations in the consistent condition should propagate a stronger signal to the RED Output node than to the GREEN Output node as a function of their respective strong activation and excitatory and inhibitory weight combinations from both HW1 and HC1. This effect however, will be attenuated in the inconsistent condition because the GREEN and RED Output nodes are receiving relatively strong inhibitory and excitatory signals from both HC2 (see top right plot), in addition to the signal from HW1. This effect is evidence of conflicting processing, and helps foreshadow the reaction time differences we’ll observe in the response time plot.

Page two of the run documentation contains two plots depicting the activation for neurons in the Output layer. As was foreshadowed in the hidden layer, the RED and GREEN nodes in the output layer have a considerable activation overlap in the inconsistent condition. In contrast, the RED Output node has a clear initial higher spike, and this activation difference continues over the course of the time spent processing. Note that although this difference looks seemingly modest as plotted from the Output layer node activations, this plays a critical role in the time course of processing and strength of activation for the actual response. This is because the evidence accumulator in the response mechanism accumulates total activation as a function of the weighted difference between these Output nodes. Thus, a consistently higher activation for the RED node at the Output layer level should lead to a steep accumulation for the response firing mechanism, which should fire as RED given the task instruction in this simulation’s input stimuli.

All of this is ultimately evidenced in page three of the run documentation, which contains a plot depicting response reaction time as a function of condition. First, one can note that the network takes approximately 250ms to respond after stimulus onset (e.g., at 0 ms). While this was not explicitly simulated by the model, this is consistent with human reaction time, which takes some time to process a visual stimulus, execute a response, etc. More interestingly, one can further note from the graph that the model responded correctly in both conditions (e.g., neither of the Green responses approached the firing threshold). Furthermore, and consistent with what is observed in human performance, one can see from the solid lines that the RED response in the consistent condition reaches the threshold and fires much faster than the RED response in the inconsistent condition. Evidence for conflict can also be gleaned from the fact that while the GREEN response quickly asymptotes to zero in the consistent condition, in the inconsistent condition the GREEN response has a strong initial spike that actually inhibits the RED response spike for approximately 200ms.

**Discussion:**

The simulations demonstrate several nuances of the human cognitive system. As evidence in the previous results section, the network was able to capture the effect of differential practice and attention allocation by adapting to training and learning to tune to task instructions. It was able to demonstrate the effects of response conflict and correctly responded to the task at hand in a time course similar to human participants. Furthermore, it accomplished this using a structure and dynamic set of parallel processing algorithms that mimic those of the human brain. However, it was not without limitations. First, the model did not explicitly simulate visual recognition, which would yield much stronger evidence for the model as an analogue to the human processing system. Furthermore, with the exception of the training phase (which used backpropagation for the weight tuning) the model was strictly feedforward. Due to time constraints, I have been unable to work out a more sophisticated algorithm to account for recurrent networks with additional inhibitory and excitatory links within layer. At this point, it is unclear whether or not this would lead to a great difference in overall performance itself, but a benefit of this would be the possibility to extend its conflict monitoring mechanisms to allow for more direct comparisons to several pieces of quantitative evidence from cognitive neuroscience. For example, the presence of such links allow one to treat the model as a Hopfield network (e.g., Botvinick et al. (2001), and calculate the Hopfield energy:

Doing so for the Output layer in this model would allow me to map this energy value across time as a function of different stimuli types, and treat it as a quantitative representation of activity in the anterior cingulate cortex (i.e., conflict). As a final note, it would also be interesting to apply this model to other response conflict tasks (e.g., base-rate problem solving). I am considering implementing parts of this model to the computational model for Bayesian reasoning in my dissertation.

References

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624-652.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, *97*(3), 332-361.

Kanne, S. M., Balota, D. A., Spieler, D. H., & Faust, M. E. (1998). Explorations of Cohen, Dunbar, and McClelland's (1990) connectionist model of Stroop performance. *Psychological review*, *105*(1), 174 - 187.

Mewhort, D. J., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 872-882.