A Production System Model of Cognitive Impairments Following Frontal Lobe Damage

Daniel Y. Kimberg
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-8117
kimberg@cmu.edu

Martha J. Farah
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-2789
farah@psy.cmu.edu

Abstract

A computer model is presented which performs four different types of tasks sometimes impaired by frontal damage: the Wisconsin Card Sorting Test, the Stroop task, a motor sequencing task and a context memory task. Patterns of performance typical of frontal-damaged patients are shown to result in each task from the same type of damage to the model, namely the weakening of associations among elements in working memory. The simulation shows how a single underlying type of damage could result in impairments on a variety of seemingly distinct tasks. Furthermore, the hypothesized damage affects the processing components that carry out the task rather than a distinct central executive responsible for coordinating these components.

Introduction

Patients with damage to the frontal cortex have difficulty with a wide range of tasks, from the execution of simple manual sequences (Luria, 1965; Kimura, 1977) to sorting stimuli into abstract categories (Milner, 1963). One of the challenges of explaining frontal function is to account for the diversity of abilities that can be impaired by frontal damage. In the present paper we attempt to capture a commonality among the failures of frontal-damaged patients in a variety of tasks. We present simulations of four different tasks at which frontal-damaged patients, particularly those with dorsolateral frontal damage, have often been found to show deficits (Stuss & Benson, 1983). They are the Stroop Task, the Wisconsin Card Sorting Test, motor sequencing tasks, and memory for context.

This research was supported by an AFOSR graduate fellowship to the first author, ONR grant N00014-91-J1546, NIMH grant R01 MH48274, NINDS career development award K04 NS01405, and a grant from the McDonnell-Pew Program in Cognitive Neuroscience to the second author, and ONR grant N00014-90-J1489 to John R. Anderson.

Stroop Task. In the Stroop task, subjects are shown color names printed in different colored inks and asked either to read the word or to name the color in which the word is printed. Normal subjects show interference when asked to name the colors of stimuli in which the color and word conflict (e.g., the word "blue" in red ink). A similar pattern of interference exists when naming the word, although the differences are much smaller. In general, frontal-damaged patients have been found to be impaired at this task, showing disproportionate interference when naming colors (Perret, 1974; Dunbar & Bub, in preparation).

Wisconsin Card Sorting Test. In the Wisconsin Card Sorting Test (WCST), patients are asked to sort a number of cards that vary according to the shape of the objects represented, the color of those objects, and the number of objects. The piles into which the cards must be sorted vary according to these same attributes, so that there is exactly one pile for each possible color, shape, and number. Initially, one of these attributes is selected as the sorting category, and the subject will be given positive feedback only if they sort the card according to that attribute. Whenever the subject sorts ten consecutive cards correctly, the category changes.

Milner (1963) found that, as compared to patients with lesions elsewhere in the brain, frontal-damaged patients made an unusually high number of perseverative errors, continuing to sort according to the previous category after the category had shifted.

Motor Sequencing. Frontal-damaged patients have also been widely documented as having difficulty with sequencing tasks, especially the sequencing of motor actions. Kolb and Milner (1981) found that among patients with a variety of lesion sites, left and right frontal-damaged patients were the most impaired at imitating sequences of facial movements, and were also impaired at imitating arm movement sequences. Similarly, Kimura (1982) found that left frontal-damaged patients, in comparison to other patient groups, were the most impaired on all forms of oral move-

ments, but especially sequences of oral movements, and that these same patients were also the most impaired on manual sequences. As well, Jason (1985) examined the performance of a variety of patient groups on a manual sequence task, and found left frontal-damaged patients to be the most impaired.

Memory for Context. While not uniformly amnesic, frontal-damaged patients often show deficits at particular memory tasks. Schacter (1987) has applied the term "spatiotemporal context" to the type of memory tasks at which frontal-damaged patients have been shown to show disproportional impairment. Parkin, Leng and Stanhope (1988) report the results of a detailed case study of a frontal-damaged patient. Among their findings, they report source amnesia, that is, impaired memory for the original context of learning, and impaired memory for temporal sequence. Janowsky, Shimamura and Squire (1989) investigated memory for recently learned facts and memory for the source of the facts in a group of frontal-damaged patients. Although the patients were normal in their ability to recall the facts, compared with age-matched control subjects, they frequently attributed the facts to incorrect sources.

These four types of task appear, on the surface at least, to be quite different from one another. Previous attempts to explain these and other frontal impairments have called into play a variety of mechanisms, including error utilization (Konow & Pribram, 1970), executive or supervisory processes (Shallice, 1982; Norman & Shallice, 1986), planning (Duncan, 1986), temporal integration of behavior (Milner, 1982), and inhibitory processes (Diamond, 1989). Here we propose a single underlying impairment that can account for the failures of frontal-damaged patients in all four of these tasks.

A production system model of the effects of frontal lobe damage

In our view, the effect of frontal lobe damage on behavior is to weaken the associations among working memory representations that include representations of goals, stimuli in the environment, and stored declarative knowledge. Thus, we hypothesize that the representation of the goals themselves is unaffected, consistent with the oft-cited observation that frontal-damaged patients can report the correct goal even while performing an inappropriate action (e.g. Konow & Pribram, 1970). We also hypothesize that the stimulus environment is perceived normally, and the full range of possible actions is available, also consistent with clinical observation. Finally, declarative knowledge is available, consistent with the results of memory research on frontal-damaged patients. We hypothesize a functional attenuation of association strengths among these different working memory representations. In effect, the differing degrees of mutual relevance among goals, stimuli and stored knowledge become less discriminable after frontal lobe damage.

We have chosen to implement our model using a simplified subset of the ACT-R framework (Anderson, 1983; 1989; in preparation). It should be noted that the architecture of this system was designed to account for normal cognition in a variety of tasks, drawing upon empirical findings on normal human learning and memory (Anderson, 1983) and upon a "rational analysis" of human cognition (Anderson, 1989). Thus, to the extent that the present model can account for the behavior of frontal-damaged patients, it does so without any ad hoc features designed specifcally for that purpose.

ACT-R is a production system, incorporating as its procedural knowledge a set of IF-THEN rules. These rules specify actions and the conditions under which they should be performed. In addition, it includes a working memory representation in which declarative knowledge is represented. The behavior of this system depends on which productions are selected for execution, according to two mechanisms: *matching* and *conflict resolution*. Matching refers to the process by which it is determined whether or not each production's conditions hold. This is accomplished by comparing the conditions of the rule to the contents of working memory. If more than one production matches the contents of working memory, as is often the case, then the process of conflict resolution is used to select a single production, as only one can be executed at one time. In the model described below, conflict resolution is accomplished by comparing activation levels, with the most active production being executed. There are four different sources that contribute to each production's activation:

Baseline activation is the invariant activation associated with a particular production. The higher the prior probability that the particular production will be applicable, the higher its baseline activation level. Productions with higher probabilities of being applicable, and therefore higher baseline activation, are more likely overall to fire.

Priming activation is additional activation that a particular production receives when it is executed. Priming activation falls off over the next several cycles of the simulation, and reflects the likelihood that a production that has just been executed will be applicable again in the very near future.

Noise activation is also present in the system.

Data activation is the activation added to a production from the working memory elements (WMEs) with which the production matches. The contribution of data activation is the sum of the activations of those WMEs. A WME's activation is calculated from its previous activation, from the previous activations of those WMEs with which it is connected, and from the strengths of those connections. However, in the present model, the previous activation of each WME is held constant. Thus, the activation of a particular

WME depends only on the strengths of its connections with other WMEs. Furthermore, since each production refers only to a subset of the entire working memory representation, only connections between pairs of WMEs both matched by that production will contribute to its activation.

Consider a production that matches both a goal WME (e.g., name the ink color of a stimulus) and some stimulus attribute WME (e.g., the ink color is red). If the goal is strongly associated with that stimulus attribute (as in this case, most likely) then the production will receive a large amount of activation from its data and will be more likely to fire. If the goal is only weakly associated with the stimulus (if the attribute WME in the above example were the lexical identity instead of the ink color) then the production will receive less activation from its data and will be less likely to fire. In this way, the model is more likely to execute a production for which mutually relevant goal and stimulus attributes are present.

Simulations of the four tasks.

In this section we present a simulation of the four tasks, and examine the effect of weakening association strengths among WMEs on the performance of the simulations. We first describe the undamaged models and then the results of damaging the normal system. Note that the same parameters (noise, priming activation, and decay rate) were used in all four simulations.

Stroop task. The Stroop task simulation consists of two productions, *name-color* and *name-word*, corresponding to the two potential responses to each stimulus. The attribute to be named for a given set of trials is set by strengthening the connection between the appropriate attribute WME (e.g. *colorname* when the goal is to name the color) and a WME which maintains information about the task (such as the current stimulus). Each production only matches against the appropriate attribute WME, so that the *name-color* production will only receive activation from the attribute WME *colorname*. The productions also receive activation from the connection between their attributes and the data they match. So the word naming production receives activation from the strong connections between the word attribute WME and word data, while the color naming task receives activation from the weaker but still strong connections between that attribute and the color data. Also, the baseline activation of the *name-word* production is stronger than that of *name-color*, consistent with the more frequent use of word naming in everyday life (see MacLeod, 1991).

At the presentation of a stimulus, both productions are placed in the conflict set, since all stimuli in this simulation have both color name and word name attributes. However, the correct task WME will receive more activation from the strengthened connection to the relevant attribute. And the discriminability will be greater when the task is word naming, since its production has a greater baseline activation.

Wisconsin Card Sorting Test. The WCST simulation consists of six productions: three for sorting and three for utilizing feedback. Each of the three sorting productions sorts by a particular attribute – color, shape, or number. Thus, whenever the production *sort-by-number* fires, the current card is sorted according to its number attribute.

The three feedback productions model how a subject should ideally utilize feedback, by constraining which categories will be sources of activation. After positive feedback, the current category is made the only category eligible to be a source of activation. After negative feedback, the incorrect category is made no longer a source of activation. And if this results in an empty set, the other two possible categories are then made eligible again. The model always implicitly knows which sorting categories are potentially correct, because only those categories are potential sources of activation. This is analogous to how, in the Stroop task, only the correct attribute (color or word name) is a strong source of activation for its corresponding production. However, the WCST uses the eligibility set to change these biases between trials.

Since the feedback mechanism just affects the eligible sources of activation, not whether or not particular WMEs are in working memory, this information does not directly constrain which productions can match. Instead, it biases which categories will be considered, by providing a source of activation for only those sorting productions whose categories are still eligible. For example, the production *sort-by-color* would receive activation from the connection between the *colorsort* WME and the list of possible categories. Since *colorsort* is most strongly associated with the *color* category, this production would be strongest when *color* was still an eligible category.

Initially, the set contains all three categories, so as to be unbiased. At present, the WCST productions do not wait for a number of correct trials before proceeding, but simply shift after a fixed interval.

When a card is presented, all three sorting productions are in the conflict set. The productions whose categories are still eligible receive activation from the connection between their categories and the corresponding task nodes. After a sort, one of the three feedback productions will match. There is one production to handle positive feedback, and two for negative feedback (when the eligibility set is larger than 1 or equal to 1). After feedback, the model attempts to sort the next card.

Motor sequencing task. The simulation of the motor sequencing task is the simplest. The model is presented with a repeating sequence of stimuli, each of

which requires a distinct response. The stimuli can be thought of as devices, and the responses as different motor actions, similar to the task used by Kimura (1977). A different production for each potential response matches against both the action to be performed and the device to be acted upon in working memory. Since matches with congruent actions and devices will benefit from strong connections, they will receive more activation from data. There are five possible devices and five corresponding actions. The sequence of action in the simulation is also straightforward: the first stimulus is presented, and all five motor productions are in the conflict set. The correct one receives more data activation and is most likely to fire. Then the second stimulus is presented.

Memory. Memory for context can be modeled using a single production, *name-context*, to name the context of a presented item. Since the same production is used to name either the correct or incorrect context, in this simulation *name-context* competes only with itself. Different *instantiations* of the production, corresponding to the different contexts, are all in the conflict set simultaneously. However, since these instantiations refer to different subsets of working memory, they receive potentially different levels of activation, and can therefore be discriminated.

Context memory is simulated using a different WME for each context in which information is presented. Each context WME is in turn associated with a set of WMEs which represent the features of the environment. The unique subset of features with which a context WME is associated defines that context. Context memory can then be seen as the ability to name the appropriate subset of features through a label for those features. While this is a oversimplification of the notion of context, it preserves the critical requirement that the model must produce some element unique to the original situation in which a test item was presented, namely a label for that particular conjunction of features. Since the acquisition process is not simulated here, only the testing phase, this requires just a single production – one to name the context of the presented item. Each item is strongly associated with the features of the context in which it originally appeared, and weakly with the other features. The production *name-context* matches all possible available contexts, but each instantiation receives data activation from the connections between a particular context's features and the probe item. Thus, the production will be more likely to fire with the correct context, since that instantiation maximizes the amount of data activation it will receive.

To model the task, a stimulus is presented along with a set of five possible contexts. Thus five instantiations of *name-context* are placed in the conflict set, one for each context. Each instantiation matches against the stimulus and against the features of its context. The instantiation in which the stimulus is most closely associated with the features of the context is the one most likely to fire.

Recognition memory is modeled here as a special case of context memory, in which the subject must decide whether or not the stimulus was originally encoded in the experimental context. In the present simulation, this requires an additional production, *fail-to-recognize*, which produces the default behavior of failing to recognize an item. Because *fail-to-recognize* is a default, it will always match on the basis of its own baseline activation. That baseline activation therefore represents a threshold for recognizing an item. When an instantiation of *name-context* exceeds this threshold, the item is in effect recognized. Otherwise, *fail-to-recognize* will fire as a default. Note that *fail-to-recognize* would probably never fire in the case of context judgements, since it is extremely unlikely that multiple contexts, would all fail to reach threshold on the same trial. In this way, one might say that context memory judgements are between two or more real contexts, while recognition memory judgements are in effect between the correct context and a default context.

The simulation was run on all four tasks, using fixed sequences of stimuli. Each simulation was first run normally, then damaged. The simulation was damaged by weakening all of the connections between working memory elements by either 50% or 80%.

Results

Errors under each condition were tabulated as either non-perseverative or perseverative, except in the memory task, for which there was only one production, and thus no possibility for perseveration in the present model. If the incorrectly fired production had been fired on the immediately preceding trial (whether or not correctly then) it was counted a perseveration. Although it is possible that there would be perseveration due to data priming, this source of activation was not included in this simulation. Also note that the gradual decay of priming activation may cause perseverations over intervening steps, although these are here counted as non-perseverative errors.

The results of the tasks are presented in Table 1.

	Normal		50%		80%	
	NP	P	NP	P	NP	P
StroopColor	9	1	21	56	38	245
StroopWord	0	0	2	0	13	67
WCST	0	4	24	58	81	282
Motor	0	0	23	54	164	367
Context	2		81		379	
Recognition	0		1		22	

Table 1: Total non-perseverative (NP) and perseverative (P) errors on each task (1000 trials for each task for each damage level).

Without damage, the model performs all of the tasks at a high level. Although the noise makes errors possible, the high discriminability of the productions in the conflict set makes errors extremely unlikely. With 50% damage, more errors are made on all tasks, and there is a clear bias towards perseverative errors. Finally, with 80% damage, there is a greater proportion of errors, and a greater proportion of those errors are perseverative.

In the Stroop task, the damaged model shows interference from the unattended attribute. Moreover, as in frontal-damaged patients, there is more interference in the color naming condition than in the word reading condition. What is responsible for this pattern of results? In the undamaged model, the discriminability of the word-naming and color-naming productions is high, due to the strong connections between the appropriate attribute WME and the WME which maintains information about the task. When the connections are weakened, however, the activations of the two productions become more similar, and noise activation is therefore more likely to cause the wrong production to be selected. The fact that color naming is more vulnerable to intrusions by word naming than vice versa is explained by the higher baseline activation of word naming, which results from its more frequent use.

In the WCST, the damaged model simulates patient behavior in perseverating sorting categories even after negative feedback. As before, the reason for this can be understood by first considering the functioning of the normal system. Normally, feedback affects the selection of a sorting category by determining which categories remain in the eligible set. This set biases the model towards eligible categories through connections with the possible sorting WMEs. While the damaged model still uses feedback to constrain the eligible categories, the weakened connections reduce the magnitude of this bias. This reduces the discriminability of the different sorting productions to the level where noise activation can sometimes cause an inappropriate production to be the most active. The perseverative character of many of the errors results from priming activation causing recently selected productions to be especially active.

In the motor sequencing task, damage causes the model to associate incorrect actions with devices. Unlike the previous two tasks, in this task frontal-damaged patients show nonperseverative as well as perseverative errors. Both types of errors are also made by the damaged model. Why does the damaged model behave in this way? While normally the correct action can be discriminated on the basis of its greater association with the current device, this connection is weakened by damage. Noise activation will cause incorrect productions to be selected, out of sequence, and priming activation will bias these errors towards perseveration.

In the memory tasks, damage leads to impaired performance at context memory, but not at recognition memory. As modeled here, discrimination of the correct context depends on the connections between the item to be recognized and the features of that context. When these connections are weakened, the presence of noise makes it more likely that a similar but less appropriate context will receive greater activation. Recognition memory is the only case, among the four tasks simulated, which does not require the discrimination of a particular response among close competitors, and thus is not especially harmed by the damage manipulation. Interestingly, the model predicts that frontal-damaged patients will be much better at recognition than at context memory, but not necessarily normal. Again, there is not sufficient published data to address this definitively. In one study (Parkin, Leng, & Stanhope, 1988) both normal and patient groups appeared to be near ceiling, while in another study (Janowsky et al., 1989), there was a non-significant trend in this direction.

Discussion

We believe that there are two general aspects of the effects of frontal damage that have made the underlying nature of frontal lobe function so elusive. First, frontal damage can affect performance on a wide variety of tasks that do not seem, on the surface, to have anything in common. Second, when frontal damage impairs task performance, it does so without impairing patients' knowledge of the task goals, their perception of the relevant stimuli, their ability to execute the individual actions, or their memory for previously learned facts. These two factors have given rise to the idea of the frontal cortex as a single "central executive," which is called into play regardless of the cognitive domain. This central executive is required when the activities of multiple components of the cognitive architecture must be coordinated, but would not necessarily be required for performing simpler tasks. The model described here, however, provides a unified account of four deficits sometimes arising from frontal lobe damage, and does so without postulating damage to a central executive, but rather to the processing components that are used to perform the task.

This explanation also highlights what is common among the tasks failed by frontal-damaged patients. In all of the tasks modeled here, several sources of information compete to guide behavior. One of these sources in particular, connections among internal representations, is critical for differentiating among several close competitors. When these connections are weakened, other sources of activation (e.g., priming and noise) become more important in determining behavior. While the damaged model might be described accurately as unable to make use of errors (as in the WCST), impulsive (failing to inhibit inappropriate responses, as in the Stroop Task), perseverative (as in the motor task), or impaired in the use of spatiotemporal context (as in the memory tasks), a single functional deficit can ac-

count for all of these deficits.

The present model might seem to imply that, in conflict with common clinical observation, all frontal-damaged patients should fail all of the tasks described here. However, there could be distinct areas in frontal cortex sharing the same abstract function, namely the maintenance of working memory associations, but differing as to the types of working memory elements represented. One would therefore not expect the same patient to show impairment on exactly these four or any particular set of tasks. This type of organization, of a large cortical area operating according to common information processing mechanisms but subdivided into distinct and dissociable modules according to the content of the information represented, can also be seen in the visual cortex. Numerous areas in the extrastriate visual cortex share common functional mechanisms (e.g., retinotopy, the integration of information from earlier visual areas, and center-surround organization), but differ in the type of visual information represented in these maps (Cowey, 1982).

Are there any tasks at which the model predicts frontal-damaged patients would be unimpaired? In fact, while the underlying deficit that we have hypothesized is quite general, it is restricted to tasks that tax the conflict resolution process, that is, the ability to select among a group of potentially relevant actions. Thus, we would predict normal performance on tasks in which: the potentially relevant actions or responses are narrowed down by a "structured" or highly constrained task or stimulus environment (in the model, few productions matching the active working memory elements); the appropriate responses are highly routinized (in the model, high baseline activation in the appropriate production); or there are pronounced differences in the relevance of the available responses (in the model, pronounced differences in the association strengths among working memory elements matched by appropriate and inappropriate productions). This accords well with the common clinical observation that frontal-damaged patients may do well on relatively structured tasks or very familiar tasks, somewhat independent of difficulty, despite failing dramatically on the types of tasks modeled here.

Acknowledgements. The authors thank John Anderson, Prahlad Gupta, Paul Reber, Bob Stowe, and Don Stuss, for valuable comments.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Cowey, A. (in press). Cortical visual areas and the neurobiology of higher visual processes. In M. J. Farah and G. Ratcliffe (Eds.), *The Neural Bases of High-Level Vision: Collected Tutorial Essays.* Hillsdale, NJ: Erlbaum.

Diamond, A. (1989). Developmental progression in human infants and infant monkeys, and the neural bases of inhibitory control of reaching. In A. Diamond (Ed.), *The development and neural bases of higher cognitive functions.* New York: NY Academy of Science Press.

Duncan, J. (1986). Disorganisation of behaviour after frontal lobe damage. *Cognitive Neuropsychology, 3,* 271-290.

Janowsky, J. S., Shimamura, A. P., & Squire, L. R. (1989). Source memory impairment in patients with frontal lobe lesions. *Neuropsychologia, 27,* 1043-1056.

Jason, G. W. (1985). Manual sequences learning after focal cortical lesions. *Neuropsychologia, 23,* 483-496.

Kimura, D. (1977). Acquisition of a motor skill after left-hemisphere damage. *Brain, 100,* 527-542.

Kimura, D. (1982). Left-hemisphere control of oral and brachial movements and their relation to communication. *Philosophical Transactions of the Royal Society of London B, 298,* 135-149.

Kolb, B. & Milner, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia, 19,* 491503.

Konow, A. & Pribram, K. H. (1970). Error recognition and utilisation produced by injury to the frontal cortex in man. *Neuropsychologia, 8,* 489-491.

L'hermitte, F. (1983). "Utilization behavior" and its relation to lesions of the frontal lobes. *Brain, 106,* 237-255.

Luria, A. R. (1965). Two kinds of motor perseveration in massive injury of the frontal lobes. *Brain, 88,* 1-10.

Luria, A. R. (1966). *Higher cortical functions in man.* London: Tavistock.

Luria, A. R. (1973). *The Working Brain.* New York: Basic Books.

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

Milner, B. (1963). Effects of different brain lesions on card sorting. *Archives of Neurology, 9,* 90-100.

Milner, B. (1982). Some cognitive effects of frontal-lobe lesions in man. *Philosophical Transactions of the Royal Society of London B, 298,* 211-226.

Norman, D. A. & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. E. Shapiro (Eds.), *Consciousness and Self-Regulation, Vol. 4.* New York: Plenum Press.

Parkin, A. J., Leng, N. R. C., & Stanhope, N. (1988). Memory impairment following ruptured aneurysm of the antreior communicating artery. *Brain and Cognition, 7,* 231-243.

Perret, E. (1974). The left frontal lobe of man and the suppression of habitual responses in verbal categorical behavior. *Neuropsychologia, 12,* 323-330.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B, 298,* 199-209.

Stuss, D. T. & Benson, D. F. (1983). Frontal lobe lesions and behavior. In A. Kertesz (Ed.), *Localization in Neuropsychology.* New York: Academic Press.