

Neuropsychological inference with an interactive brain: A critique of the “locality” assumption

Martha J. Farah

Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6196

Electronic mail: *mfarah@cattell.psych.upenn.edu*

Abstract: When cognitive neuropsychologists make inferences about the functional architecture of the normal mind from selective cognitive impairments they generally assume that the effects of brain damage are local, that is, that the nondamaged components of the architecture continue to function as they did before the damage. This assumption follows from the view that the components of the functional architecture are modular, in the sense of being informationally encapsulated. In this target article it is argued that this “locality” assumption is probably not correct in general. Inferences about the functional architecture can nevertheless be made from neuropsychological data with an alternative set of assumptions, according to which human information processing is graded, distributed, and interactive. These claims are supported by three examples of neuropsychological dissociations and a comparison of the inferences obtained from these impairments with and without the locality assumption. The three dissociations are: selective impairments in knowledge of living things, disengagement of visual attention, and overt face recognition. In all three cases, the neuropsychological phenomena lead to more plausible inferences about the normal functional architecture when the locality assumption is abandoned. Also discussed are the relations between the locality assumption in neuropsychology and broader issues, including Fodor’s modularity hypothesis and the choice between top-down and bottom-up research approaches.

Keywords: brain lesions; cognitive architecture; face recognition; localization; modularity; neural nets; neuropsychology; semantics; vision

The fact that the various parts of the encephalon, though anatomically distinct, are yet so intimately combined and related as to form a complex whole, makes it natural to suppose that lesions of greater or lesser extent in any one part should produce such general perturbation of the functions of the organ as a whole as to render it at least highly difficult to trace any uncomplicated connection between the symptoms produced and the lesion as such.

Ferrier (1886)

1. Introduction

Brain damage often has rather selective effects on cognitive functioning, impairing some abilities while sparing others. Psychologists interested in describing the “functional architecture” of the mind, that is, the set of relatively independent information-processing subsystems that underlies human intelligence, have recognized that patterns of cognitive deficit and sparing after brain damage are a potentially useful source of constraints on the functional architecture. In this target article I wish to focus on one of the assumptions that frequently underlies the use of neuropsychological data in the development of cognitive theories.

1.1. The locality assumption

Cognitive neuropsychologists generally assume that damage to one component of the functional architecture will have exclusively “local” effects. In other words, the non-damaged components will continue to function normally and the patient’s behavior will therefore manifest the underlying impairment in a relatively direct and straightforward way. This assumption follows from a view of the cognitive architecture as “modular,” in the sense of being “informationally encapsulated” (Fodor 1983; see also multiple book review, *BBS* 8(1)1985).

According to this version of the modularity hypothesis, the different components of the functional architecture do not interact with one another except when one has completed its processing, at which point it makes the end product available to other components. Even these interactions are limited, so that a given component receives input from relatively few (perhaps just one) of the other components. Thus, a paradigm module takes its input from just one other component of the functional architecture (e.g., phonetic analysis would be hypothesized to take its input just from prephonetic acoustic analysis), carries out its computations without being affected by other information available in other components (even potentially relevant information, such as semantic context), and then presents its output to the next component in line, for which it might be the sole input (e.g., the auditory input lexicon, which

would again be hypothesized to take only phonetic input).

In such an architecture, each component minds its own business and knows nothing about most of the other components. What follows for a damaged system is that most of the components will be oblivious to the loss of any one, carrying on precisely as before. If the components of the functional architecture were informationally encapsulated then the locality assumption would hold; the removal of one component would have only very local effects on the functioning of the system as a whole, affecting performance only in those tasks that directly call upon the damaged component. Indeed, one of Fodor's other criteria for modulehood, which he suggests will coincide with informational encapsulation, is that modules make use of dedicated hardware and can therefore be selectively impaired by local brain damage. In contrast, if the different components of the cognitive system were highly interactive, each one depending on input from many or most of the others, then damage to any one component could significantly modify the functioning of the others.

Several cognitive neuropsychologists have pointed out that informational encapsulation and the locality of the effects of brain damage are assumptions, and they have expressed varying degrees of confidence in them (Allport 1985; Caplan 1981; Humphreys & Riddoch 1987; Kinsbourne 1971; Kosslyn & Van Kleek 1990; Moscovitch & Umiltà 1990; Shallice 1988; von Klein 1977). For example, Shallice (1988, Ch. 2) endorses a weaker and more general version of modularity than Fodor's, according to which components of the functional architecture can be distinguished, (1) conceptually in terms of their specialized functions and (2) empirically by the relatively selective deficits that ensue upon damage to one of them. He likens this concept of modularity to Posner's (1978) "isolable subsystems" and offers the following criterion from Tulving (1983) for distinguishing modular systems with some mutual dependence among modules from fully interactive systems: components of a modular system in this weaker sense may not operate as efficiently when other components have been damaged but they will nevertheless continue to function roughly normally. According to this view, the locality assumption is not strictly true, but it is nevertheless roughly true: one would not expect *pronounced* changes in the functioning of nondamaged components.

Closely related to the locality assumption is the "transparency assumption" of Caramazza (1984; 1986). Although different statements of this assumption leave room for different interpretations, it is probably weaker than the locality assumption. Particularly in more recent statements (e.g., Caramazza 1992), it appears transparency requires only that the behavior of the damaged system be *understandable* in terms of the functional architecture of the normal system. Changes in the functioning of nondamaged components are not considered a violation of the transparency assumption so long as they are understandable. In particular, interactivity and consequent nonlocal effects are permitted; presumably only if the nonlocal interactions became unstable and chaotic would the transparency assumption be violated.

Unlike the weaker transparency assumption, the locality assumption licenses quite direct inferences from the manifest behavioral deficit to the identity of the underlying damaged cognitive component, inferences of the form "selective deficit in ability A implies a compo-

nent of the functional architecture dedicated to A." Obviously such inferences can go awry if the selectivity of the deficit is not real, for example, if the tasks testing A are merely harder than the comparison tasks, if there are other abilities that are not tested but are also impaired, or if a combination of functional lesions is mistaken for a single one (see Shallice 1988, Ch. 10, for a thorough discussion of other possibilities for misinterpreting dissociations in a weakly modular theoretical framework). In addition, even simple tasks tap several components at once, and properly designed control tasks are needed to pinpoint the deficient component and absolve intact components downstream. However, assuming that the relevant ability has been experimentally isolated and that the deficit is truly selective, the locality assumption allows us to delineate and characterize the components of the functional architecture in a direct, almost algorithmic way.¹

1.2. The locality assumption is ubiquitous in cognitive neuropsychology

At this point the reader may think that the locality assumption is naive and that the direct inferences that it licenses constitute a mindless reification of deficits as components of the cognitive architecture, something "good" cognitive neuropsychologists would not do. Note, however, that the locality assumption is justifiable in terms of informational encapsulation. Furthermore, whether or not this seems an adequate justification, many of the best-known findings in neuropsychology fit this form of inference. A few examples will be given here and three more will be discussed in detail later (perusal of recent journals and textbooks in cognitive neuropsychology will reveal many more examples).

With the domain of reading, phonological dyslexics show a selective deficit in tasks that require grapheme-to-phoneme translation; they are able to read real words (which can be read by recognizing the word as a whole), they can copy and repeat nonwords (demonstrating intact graphemic and phonemic representation), but they cannot read nonwords, which must be read by grapheme-to-phoneme translation. This has been interpreted as an impairment in a grapheme-to-phoneme translation mechanism and hence as evidence for the existence of such a mechanism in the normal architecture (e.g., Coltheart 1985). Similarly, in surface dyslexia a selective deficit in reading irregular words with preserved regular word and nonword reading has been used to identify a deficit in whole-word recognition and hence to infer a whole-word reading mechanism distinct from the grapheme-to-phoneme route (e.g., Coltheart 1985).

In the production and understanding of spoken language, some patients are selectively impaired in processing closed class, or "function" words, leading to the conclusion that these lexical items are represented by a separate system, distinct from open class or "content" words (e.g., Zurif 1980).

In the domain of vision, some right hemisphere-damaged patients show an apparently selective impairment in the recognition of objects viewed from unusual perspectives. This has been taken to imply the existence of a stage or stages of visual information processing concerned specifically with shape constancy (e.g., Warrington 1985). Highly selective deficits in face recognition

have been taken to support the existence of a specialized module for face recognition, distinct from more general-purpose recognition mechanisms (e.g., DeRenzi 1986).

In the domain of memory, the finding that patients can be severely impaired in learning facts and other so-called declarative or explicit knowledge while displaying normal learning of skills and other forms of implicit knowledge is interpreted as evidence for multiple learning systems, one of which is dedicated to the acquisition of declarative knowledge (e.g., Squire 1992).

Some of these inferences may well be proved wrong in the light of further research. For example, perhaps there is a confounding between the factor of interest and the true determinant of the deficit. In the case of aphasics who seem selectively impaired at processing closed class words, perhaps speech stress pattern, and not lexical class, determines the boundaries of the deficit. Critical thinkers may find reasons to question the inferences in any or all of the examples given above. However, note that in most cases the question will concern the empirical specifics of the case, such as stress pattern versus lexical class. In the course of scientific debate on these and other deficits, the *form* of the inference is rarely questioned. If we can truly establish a selective deficit in ability *A* then it seems reasonable to attribute the deficit to a lesion of some component of the functional architecture that is dedicated to *A*, that is, necessary for *A* and necessary only for *A*. We are, of course, thereby assuming that the effects of the lesion on the functioning of the system are local to the lesioned component.

1.3. Two empirical issues about the locality assumption

Although it is reasonable to assume that the effects of a lesion are confined to the operation of the lesioned components and the relatively small number of components downstream in a system with informationally encapsulated modules, we do not yet know whether the brain is such a system. There is, in fact, some independent reason to believe it is not. Neurologists have long noted the highly interactive nature of brain organization and the consequent tendency for local damage to unleash new emergent organizations or modes of functioning in the remaining system (e.g., Ferrier 1886; Jackson 1873). Of course, the observations that led to these conclusions were not primarily of cognitive disorders. Therefore, whether or not the locality assumption holds in the domain of cognitive impairments, at least to a good approximation, is an open empirical question.

Note that we should be concerned more about "good approximations" than precise generalizations to neuropsychological methodology. As already mentioned, Shallice (1988) has pointed out that modularity versus interactionism is a matter of degree. From the point of view of neuropsychological methodology, if nonlocal interactions were to modulate weakly the behavior of patients after brain damage, this would not necessarily lead to wrong inferences using the locality assumption. In such a case, in which the remaining parts of the system act ever-so-slightly differently following damage, the cognitive neuropsychologist would simply fail to account for 100% of the variance in the data (not a novel experience for most of us) but would make the correct inference about functional architecture. If deviations from locality were a first-order

effect, however, then the best-fitting theory for the data using the locality assumption would be false.

There is a second question concerning the locality assumption: Is it really indispensable to cognitive neuropsychology? Must we abandon all hope of relating patient behavior to theories of the normal functional architecture if lesions in one part of the system can change the functioning of other parts? Like the first question, this one too is a matter of empirical truth or falsehood.

Nevertheless, unlike many empirical questions, these two are not of the type that lend themselves to single critical experiments. They concern very general properties of the functional architecture of cognition and our ability to make scientific inferences about complex systems using all the formal and informal methods and types of evidence available to us. The most fruitful approach to answering these two questions would therefore involve an analysis of the body of cognitive neuropsychological research, or at least an extensive sample of it.

As a starting point, I will describe three different neuropsychological dissociations that have been used to make inferences about the functional architecture of the mind. The aspect of cognition under investigation in each case is different: semantic memory, visual attention, and the relation between visual recognition and awareness. What all three have in common is the use of the locality assumption. For each I will explore alternative inferences about the functional architecture that are not constrained by the locality assumption.

How will such explorations answer the questions posed above? We can assess the empirical basis for the locality assumption by comparing the conclusions about functional architecture that are arrived at with and without it. Specifically, we can determine which conclusions are preferable, in the sense of being simpler and according better with other, independent evidence about the functional architecture. If the locality assumption generally leads to preferable conclusions, this suggests that we are probably justified in using it. However, if it often leads to nonpreferable conclusions, this suggests we should not assume that the effects of brain damage on the functioning of the cognitive architecture are local. The question of whether it is possible to draw inferences about the functional architecture from neuropsychological dissociations without the locality assumption will also be addressed by the degree to which sensible conclusions can be reached without it.

1.4. An architecture for interactive processing

Of course, comparisons between the results of inferences made with and without the locality assumption will be meaningful only if both types of inferences are constrained in principled ways. The locality assumption is one type of constraint on the kinds of functional architectures that can be inferred from a neuropsychological dissociation. It limits the elements in our explanation of a given neuropsychological deficit to just those in the normal functional architecture (minus the damaged component), operating in their normal fashion. If we simply eliminate that constraint without replacing it with other principled constraints on how local damage affects the remaining parts of the system then the comparison proposed above will not be fair to the locality assumption. We

could, of course, pick the simplest, most appealing model of the normal functional architecture and say "the way in which the remaining parts of the system change their functioning after damage produces this deficit," without saying why we chose to hypothesize *that* particular change in functioning as opposed to some other that cannot explain the deficit.

The parallel distributed processing (PDP) framework will be used as a source of principled constraints on the ways in which the remaining parts of the system behave after local damage. Computer simulation will be used to test the sufficiency of the PDP hypotheses to account for the dissociations in question. Readers who would like a detailed introduction to PDP are referred to Rumelhart and McClelland's (1986) collection of readings. For present purposes, the relevant principles of PDP are:

Distributed representation of knowledge. In PDP systems, representations consist of patterns of activation distributed over a population of units. Different entities can therefore be represented using the same set of units, because the pattern of activation over the units will be distinctive. Long-term memory knowledge is encoded in the pattern of connection strengths distributed among a population of units.

Graded nature of information processing. In PDP systems processing is not all or none: representations can be partially active, for example, through partial or sub-threshold activation of some of those units that would normally be active. Partial knowledge can be embodied in connection strengths, either before learning has been completed or after partial damage.

Interactivity. The units in PDP models are highly interconnected and thus mutual influence among different parts of the system is the rule rather than the exception. This influence can be excitatory, as when one part of a distributed representation activates the remaining parts (pattern completion), or it can be inhibitory, as when different representations compete with one another to become active or to maintain their activation. Note that interactivity is the aspect of the PDP framework that is most directly incompatible with the locality assumption. If the normal operation of a given part of the system depends on the influence of some other part, it may not operate normally after that other part has been damaged.

The psychological plausibility of PDP is controversial but it need not be definitively established here before proceeding. Instead, just as locality is being identified as an assumption and evaluated, so PDP is to be evaluated as a specific alternative assumption. In addition, as will be discussed further in the "General Discussion" (sect. 3), much of the controversy surrounding PDP concerns its adequacy for language and reasoning. It is possible that the arguments advanced here will not generalize to these cognitive domains.

2. Reinterpreting dissociations without the locality assumption: Three case studies

2.1. The functional architecture of semantic memory: Category-specific?

The existence of patients with apparent category-specific impairments in semantic memory knowledge has led to the inference that semantic memory has a categorical

organization, with different components dedicated to representing knowledge from different categories. The best-documented forms of category-specific knowledge deficit (as opposed to pure naming or visual recognition deficits) are the deficits in knowledge of living and nonliving things.

2.1.1. Evidence for selective impairments in knowledge of living and nonliving things. Beginning in the 1980s, Warrington and her colleagues began to report the existence of patients with selective impairments in knowledge of either living or nonliving things (Warrington & McCarthy 1983; 1987; Warrington & Shallice 1984). Warrington and Shallice (1984) described four patients who were much worse at identifying living things (animals, plants) than nonliving things (inanimate objects); all four had recovered from herpes encephalitis and had sustained bilateral temporal lobe damage. Two of the patients were studied in detail and showed a selective impairment for living things across a range of tasks, both visual and verbal. Table 1 shows examples of their performance in a visual identification task (in which they were to identify by name or description the item shown in a colored picture) and in a verbal definition task (in which the names of these same items were presented auditorially and they were to define them). Examples of their definitions are shown in Table 2. Other cases of selective impairment in knowledge of living things include additional postencephalitic patients described by Pietrini et al. (1988), Sartori and Job (1988), and Silveri and Gianotti (1988), a patient with encephalitis and strokes described by Newcombe et al. (in press), two head injury patients described by Farah et al. (1991), and a patient with a focal degenerative disease described by Basso et al. (1988). In all these cases there was damage to the temporal regions, known to be bilateral except in Pietrini et al.'s case 1 and the case of Basso et al., where there was evidence only of left temporal damage.

The opposite dissociation, namely, impaired knowledge of nonliving things with relatively preserved knowledge of living things, has also been observed. Warrington and McCarthy (1983; 1987) described two cases of global dysphasia following large left-hemisphere strokes in which semantic knowledge was tested in a series of matching tasks. Table 3 shows the results of a matching task in which the subjects were asked to point to the picture in an array that corresponded to a spoken word.

Table 1. An impairment in knowledge of living things: Performance on two tasks assessing knowledge of living and nonliving things

Case	Task	
	Picture identification	
	Living (%)	Nonliving (%)
JBR	6	90
SBY	0	75
Spoken word definition		
	Living (%)	Nonliving (%)
JBR	8	79
SBY	0	52

Table 2. *Examples of definitions of living and nonliving things*

Case	Definition
Living Things	
JBR	Parrot: don't know
	Daffodil: plant
	Snail: an insect animal
	Eel: not well
SBY	Ostrich: unusual
	Duck: an animal
	Wasp: bird that flies
	Crocus: rubbish material
	Holly: what you drink
	Spider: a person looking for things, he was a spider for his nation or country
Nonliving things	
JBR	Tent: temporary outhouse, living home
	Briefcase: small case used by students to carry papers
	Compass: tools for telling direction you are going
	Torch: hand-held light
	Dustbin: bin for putting rubbish in
SBY	Wheelbarrow: object used by people to take material about
	Towel: material used to dry people
	Pram: used to carry people, with wheels and a thing to sit on
	Submarine: ship that goes underneath the sea
	Umbrella: object used to protect you from water that comes

Their performance with animals and flowers was more reliable than with nonliving things. One subject was also tested with a completely nonverbal matching task in which different-looking depictions of objects or animals were to be matched to one another in an array; the same selective preservation of knowledge of animals relative to inanimate objects was found.

Although these patients are not entirely normal in their knowledge of the relatively spared category, they are markedly worse at recognizing, defining, or answering questions about items from the impaired category. The

Table 3. *An impairment in knowledge of nonliving things: Performance on two tasks assessing knowledge of living and nonliving things*

Case	Task	Spoken word/picture matching		
		Animals (%)	Flowers (%)	Objects (%)
VER		86	96	63
YOT		86	86	67
		Picture/picture matching		
		Animals (%)	Objects (%)	
YOT		100	69	

existence of a double dissociation makes it unlikely that a sheer difference in difficulty underlies the apparent selectivity of the deficits; some of the studies cited above tested several alternative explanations of the impairments in terms of factors other than semantic category (such as name frequency, familiarity, etc.) and failed to support them.

2.1.2. Interpretation of "living things" and "nonliving things" deficits relative to the functional architecture of semantic memory. Using the locality assumption, the most straightforward interpretation of the double dissociation between knowledge of living and nonliving things is that they are represented by two separate category-specific components of the functional architecture of semantic memory. A related interpretation is that semantic memory is represented using semantic features such as "animate," "domestic," and so on, and that the dissociations described here result from damage to these features (Hillis & Caramazza 1991). In either case, the dissociations seem to imply a functional architecture for semantic memory that is organized along rather abstract semantic or taxonomic lines. Figure 1 represents a category-specific model of semantic memory and its relation to visual perception and language.

Warrington and colleagues, however, have suggested an alternative interpretation, according to which semantic memory is fundamentally modality-specific. They argue that selective deficits in knowledge of living and nonliving things may reflect the differential weighting of information from different sensorimotor channels in representing knowledge about these two categories. They have pointed out that living things are distinguished primarily by their sensory attributes, whereas nonliving things are distinguished primarily by their functional attributes. For example, our knowledge of an animal such as a leopard, by which we distinguish it from other similar creatures, is predominantly visual. In contrast, our knowledge of a desk, by which we distinguish it from other furniture, is predominantly functional (i.e., what it is used for). Thus, the distinctions between impaired and preserved knowledge in the cases reviewed earlier may not be living/nonliving distinctions per se but sensory/functional distinctions, as illustrated in Figure 2.

The modality-specific hypothesis seems preferable to a

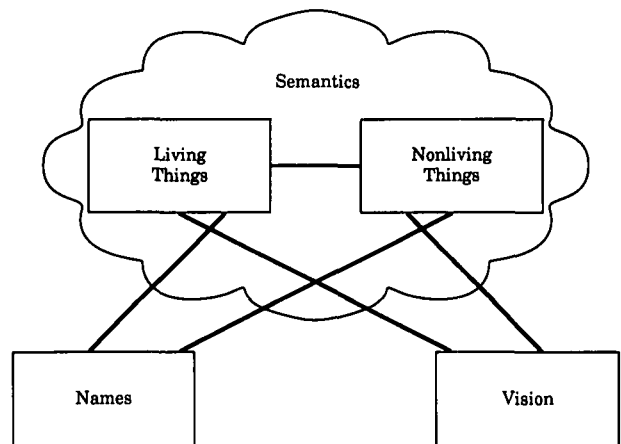


Figure 1. *Category-specific functional architecture for semantic memory.*

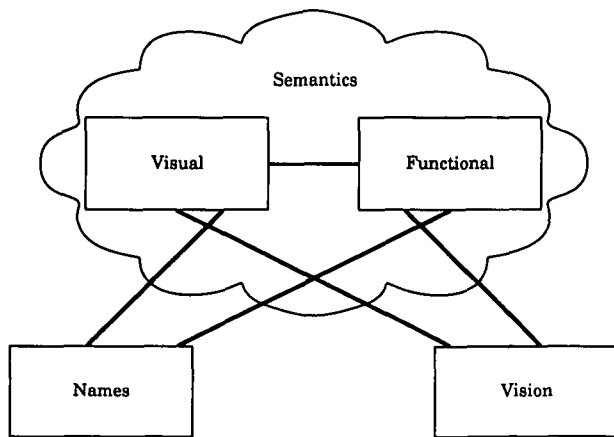


Figure 2. Modality-specific functional architecture for semantic memory.

strict semantic hypothesis for two reasons. First, it is more consistent with what is already known about brain organization. It is well known that different brain areas are dedicated to representing information from specific sensory and motor channels. Functional knowledge could conceivably be tied to the motor system. A second reason for preferring the sensory/functional hypothesis to the living/nonliving hypothesis is that exceptions to the latter have been observed in certain cases. For example, Warrington and Shallice (1984) report that their patients, who were deficient in their knowledge of living things, also had impaired knowledge of gemstones and fabrics. Warrington and McCarthy's (1987) patient, whose knowledge of most nonliving things was impaired, seemed to have retained good knowledge of very large outdoor objects such as bridges or windmills. It is at least possible that our knowledge of these aberrant categories of nonliving things is primarily visual.

Unfortunately, there appears to be a problem with the hypothesis that "living-thing impairments" are just impairments in sensory knowledge, and "nonliving-thing impairments" are just impairments in functional knowledge. This hypothesis seems to predict that cases of living-thing impairment should show good knowledge of the functional attributes of living things and cases of nonliving-thing impairment should show good knowledge of the visual attributes of nonliving things. The evidence available in cases of nonliving-thing impairment is limited to performance in matching-to-sample tasks, which does not allow us to distinguish knowledge of visual or sensory attributes from knowledge of functional attributes. However, there does appear to be adequate evidence in cases of living-thing impairment, and in at least some cases it disconfirms these predictions (for review see Farah & McClelland 1991). For example, although the definitions of living things shown in Table 2 contain little visual detail, in keeping with the sensory/functional hypothesis, they are also skimpy on functional information. If these cases had lost just their visual semantic knowledge, then why could they not retrieve functional attributes of living things, for example, the fact that parrots are kept as pets and can talk, that daffodils are a spring flower, and so on? A more direct and striking demonstration of the apparently categorical nature of the impairment is provided by

Newcombe et al. (in press), whose subject was impaired relative to normal subjects in his ability to sort living things according to such nonsensory attributes as whether or not they were generally found in the United Kingdom, in contrast to his normal performance when the task involved nonliving things.

In sum, the sensory/functional hypothesis seems preferable to the living/nonliving hypothesis because it is more in keeping with what we already know about brain organization. However, it is not able to account for the impaired ability of these patients to retrieve nonvisual information about living things.

2.1.3. Accounting for category-specific impairments with an interactive modality-specific architecture. Jay McClelland and I have modeled the double dissociation between knowledge of living and nonliving things using a simple autoassociative memory architecture with modality-specific components (Farah & McClelland 1991). We found that a two-component semantic memory system, consisting of visual and functional components, could be lesioned to produce selective impairments in knowledge of living and nonliving things. More important, we found that such a model could account for the impairment of both visual and *functional* knowledge of living things.

The basic architecture of the model is shown in Figure 2. There are three pools of units, representing the names of items, the perceived appearances of items, and the semantic memory representations of items. The semantic memory pool is subdivided into visual semantic memory and functional semantic memory. An item, living or nonliving, is represented by a pattern of +1 and -1 activations over the name and visual units, and a pattern of +1 and -1 activations over a *subset* of the semantic units. The relative proportion of visual and functional information comprising the semantic memory representation of living and nonliving things was derived empirically. Normal subjects identified terms in dictionary definitions of the living and nonliving items used by Warrington and Shallice (1984) as referring to either visual or functional properties. This experiment confirmed that visual and functional information was differentially weighted in the definitions of living and nonliving things and the results were used to determine the average proportions of visual and functional units in semantic memory representations of living and nonliving items. For the living items, about seven times as many visual semantic units than functional ones participated in the semantic memory pattern; for nonliving items the proportions were closer to equal. Units of semantic memory not involved in a particular item's representation took the activation value of 0.

The model was trained using the delta rule (Rumelhart et al. 1986) to associate the correct semantic and name portions of its pattern when presented with the visual portion as input, and the correct semantic and visual portions when presented with the name portion as input. It was then damaged by eliminating different proportions of functional or visual semantic units and its performance was assessed in a simulated picture-name matching task. In this task, each item's visual input representation is presented to the network and the pattern activated in the name units is assessed, or each pattern's name is pre-

sented and the resultant visual pattern is assessed. The resultant pattern is scored as correct if it is more similar to the correct pattern than to any of the other 19 patterns.

Figure 3A shows the averaged picture-to-name and name-to-picture performance of the model for living and nonliving items under varying degrees of damage to visual semantics. With increased damage, the model's performance drops, and it drops more precipitously for living things, in effect showing an impairment for living things comparable in selectivity to that of the patients in the literature. Figure 3B shows that the opposite dissociation is obtained when functional semantics is damaged.

The critical challenge for a modality-specific model of semantic memory is to explain how damage could create an impairment in knowledge of living things that includes

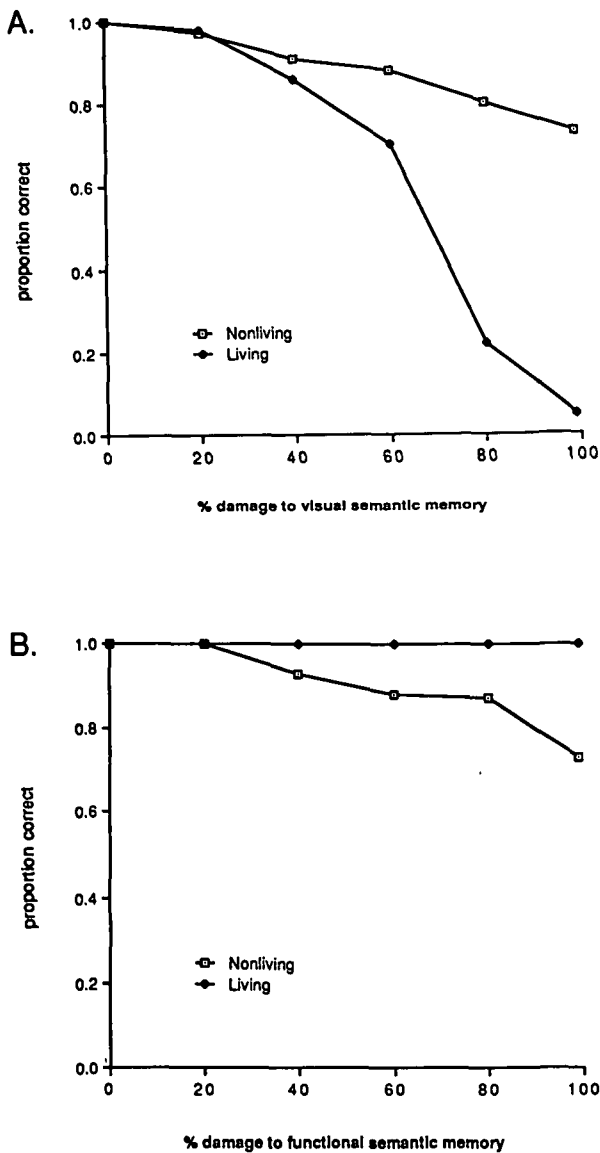


Figure 3. A: Effects of different degrees of damage to visual semantic memory units on ability of network to associate names and pictures of living things (diamonds) and nonliving things (squares). B: Effects of different degrees of damage to functional semantic memory units on ability of network to associate names and pictures of living things (diamonds) and nonliving things (squares).

functional knowledge of living things. To evaluate the model's ability to access functional semantic knowledge, we presented either name or visual input patterns as before, but instead of assessing the match between the resulting output pattern and the correct output pattern, we assessed the match between the resulting pattern in functional semantics and the correct pattern in functional semantics. The normalized dot product of these two patterns, which provides a measure between 0 (completely dissimilar) and 1 (identical), served as the dependent measure.

Figure 4 shows the accuracy with which functional semantic memory information could be activated for living and nonliving things after different degrees of damage to visual semantics. At all levels of damage, the ability to retrieve functional semantic knowledge of living things is disproportionately impaired.

These dissociations can be understood as follows. In the case of picture-name matching, the ability of a given output unit (e.g., a name unit, in the case of picture-to-name matching) to attain its correct activation value depends on the input it receives from the units to which it is connected. These consist of other name units (collateral connections) and both visual and functional semantic units. Hence the more semantic units that have been eliminated, the more the output units are deprived of the incoming activation they need to attain their correct activation values. Because most of the semantic input to the name units of living things is from visual semantics, whereas the same is not true for nonliving things, damage to visual semantics will eliminate a greater portion of the activation needed to retrieve the name patterns for living things than nonliving things, and will therefore have a more severe impact on performance.

The same principle applies to the task of activating functional semantics, although in this case the units are being deprived of collateral activation from other semantic units. Thus, when visual semantic units are destroyed, one of the sources of input to the functional semantic units

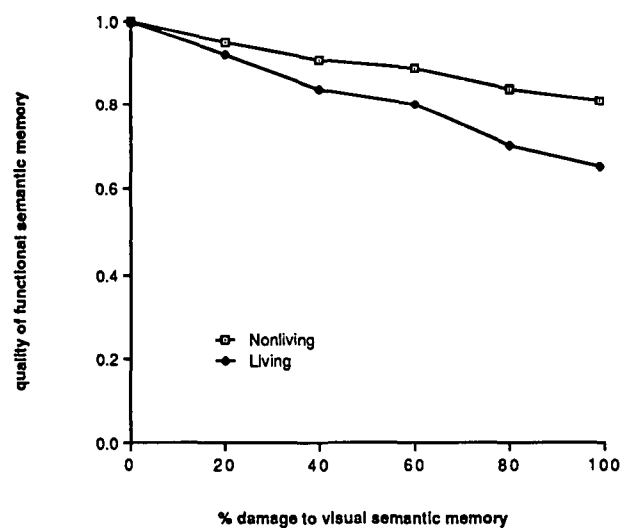


Figure 4. Effects of different degrees of damage to visual semantic memory units on ability of network to activate correct pattern in functional semantic memory units for living things (diamonds) and nonliving things (squares).

is eliminated. For living things, visual semantics comprises a proportionately larger source of input to functional semantic units than for nonliving things, hence the larger effect for these items.

2.1.4. Relevance of the locality assumption for architecture of semantic memory. Contrary to the locality assumption, when visual semantics is damaged the remaining parts of the system do not continue to function as before. In particular, functional semantics, which is part of the nondamaged residual system, becomes impaired in its ability to achieve the correct patterns of activation when given input from vision or language. This is because of the loss of collateral support from visual semantics. The ability of this model to account for the impairment in accessing functional knowledge of living things depends critically upon this nonlocal aspect of its response to damage.

2.2. The functional architecture of visual attention: A "disengage" module?

One of the best-known findings in cognitive neuropsychology concerns the "disengage" deficit that follows unilateral parietal damage. In an elegant series of studies, Posner and his colleagues have shown that parietally damaged patients have a selective impairment in their ability to disengage attention from a location in the spared ipsilesional hemifield in order to move it to a location in the affected contralesional hemifield (e.g., Posner et al. 1984). From this they have inferred the existence of a disengage component in the functional architecture of visual attention.

2.2.1. Evidence for the disengage deficit. Posner and colleagues inferred the existence of a disengage operation from experiments using a cued simple reaction time task. The typical task consists of a display, as shown in Figure 5A, which the subject fixates centrally, and in which both "cues" and "targets" are presented. The cue is usually the brightening of one of the boxes, as depicted in Figure 5B. This causes attention to be allocated to the region of space around the bright box. The target, usually a simple character such as an asterisk, is then presented in one of

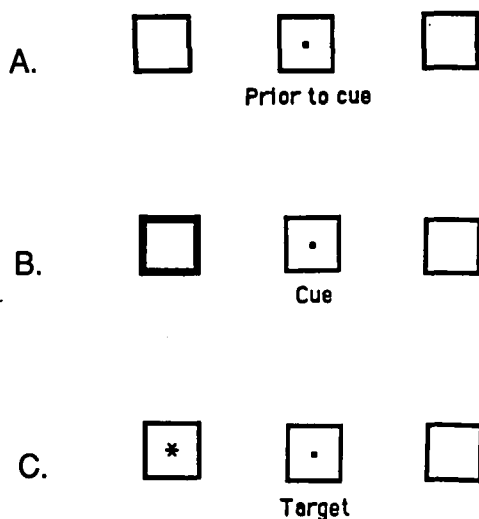


Figure 5. Sequence of trial events in the lateralized simple reaction time task: (A) fixation display; (B) cue; (C) target.

the boxes, as shown in Figure 5C. The subject's task is to press a button as soon as possible after the appearance of the target, regardless of its location. When the target is "validly" cued, that is, when it occurs on the same side of the display as the cue, reaction times to it are faster than with no cue, because attention is already optimally allocated for perceiving the target. When the target is "invalidly" cued, reaction times are slower than with no cue because attention is focused on the wrong side of space.

When parietally damaged patients are tested in this paradigm, they perform roughly normally on validly cued trials when the target appears on the side of space ipsilateral to their lesion. However, their reaction times are greatly slowed to invalidly cued contralesional targets. It is as if once attention has been engaged on the ipsilesional, or "good," side it cannot be disengaged to be moved to a target occurring on the contralesional, or "bad," side.

2.2.2. Interpretation of the disengage deficit relative to the functional architecture of visual attention. The disproportionate difficulty that parietally damaged patients have in disengaging their attention from the good side to move it to the bad side has led Posner and colleagues to infer the existence of a separate component of the functional architecture of disengaging attention. The resulting model of attention therefore postulates distinct components for engaging and disengaging attention, as shown in Figure 6.

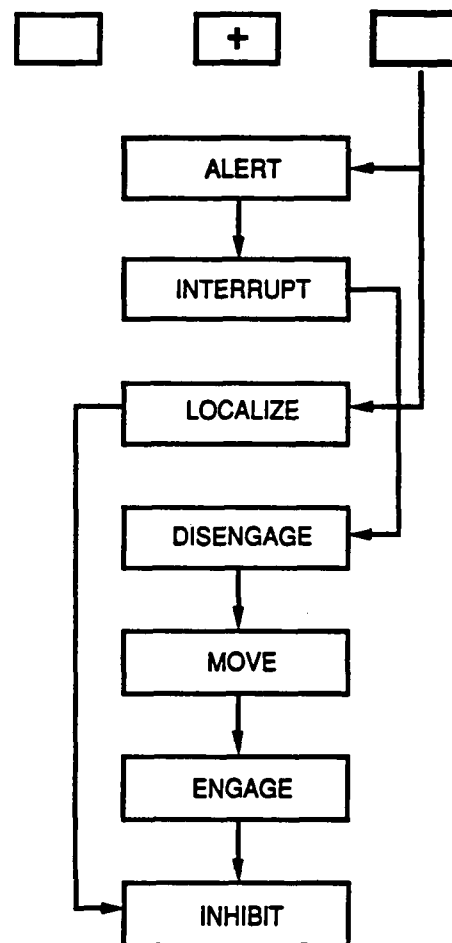


Figure 6. Functional architecture of visual attention system derived by Posner et al. (1984) from the study of brain-damaged patients.

2.2.3. Accounting for the disengage deficit with an interactive architecture that has no “disengage” component. Jonathan Cohen, Richard Romero, and I (Cohen et al., in press) have modeled normal cuing effects and the disengage deficit using a simple model of visual attention that contains no “disengage” component.

The model is depicted in Figure 7. The first layer consists of visual transducer, or input, units, through which stimuli are presented to the network. These units send their output to visual perception units, which represent the visual percept of a stimulus at a particular location in space. In this simple model there are only two locations in visual space. The visual perception units are connected to two other kinds of units. One is the response unit, which issues the detection response when it has gathered sufficient activation from the perception units to reach its threshold. We will interpret the number of processing cycles that intervene between the presentation of a target to one of the visual transducer units and the attainment of threshold activation in the response unit as a direct correlate of reaction time.

The visual perception units are also connected to a set of spatial attention units corresponding to their spatial location. The spatial units are activated by the visual unit at the corresponding location and reciprocally activate that same unit, creating a resonance that reinforces its activation. These reciprocal connections are what allow the spatial attention units to facilitate perception.

The spatial attention units are also connected to each other. For units corresponding to a given location, these connections are excitatory, that is, they reinforce each other's activation. The connections between units corresponding to different locations are inhibitory. In other words, if the units at one location are more active, they will drive down the activation of the other location's units. These mutually inhibitory connections are what give rise to attentional limitations in the model, that is, the tendency to attend to just one location at a time.

Connection strengths in this model were set by hand. Units in the model can take on activation values between 0 and 1, have a resting value of 0.1, and do not pass on activation to other units until their activation reaches a threshold of 0.9.

Before the onset of a trial, all units are at resting level

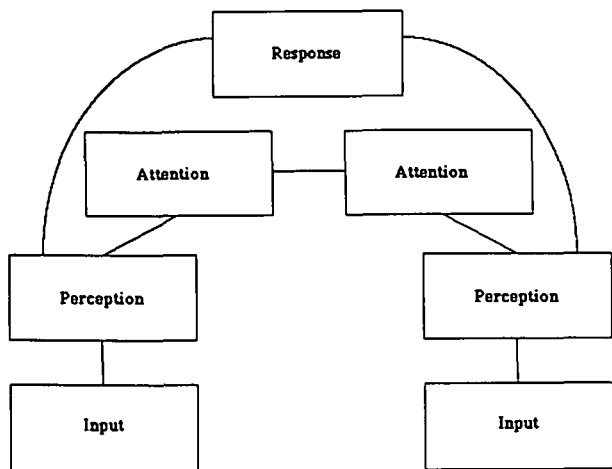


Figure 7. Functional architecture of visual attention system as modeled by Cohen et al. (in press).

activation except for the attention units, which are set to 0.5 to simulate the subject's allocation of some attention to each of the two possible stimulus locations. The presentation of a cue is simulated by clamping the activation value of one of the visual input units to 1 for the duration of the cuing interval. Presentation of the target is then simulated by clamping the activation value of one of the visual input units to 1. The target is validly cued if the same input unit is activated by both cue and target and invalidly cued if different input units are activated. We also simulated a neutral cuing condition in which no cue preceded the target. The number of processing cycles needed for the perception unit to raise the activation value of the response unit to threshold after target onset is the measure of reaction time. By regressing these numbers of cycles onto the data from normal subjects, we were able to fit the empirical data with our model.

Figure 8 shows the data from normal subjects obtained by Posner et al. (1984) and the model's best fit to the data. Why does our model show effects of valid and invalid cuing? In our model, attentional facilitation due to valid cuing is the result of both residual activation from the

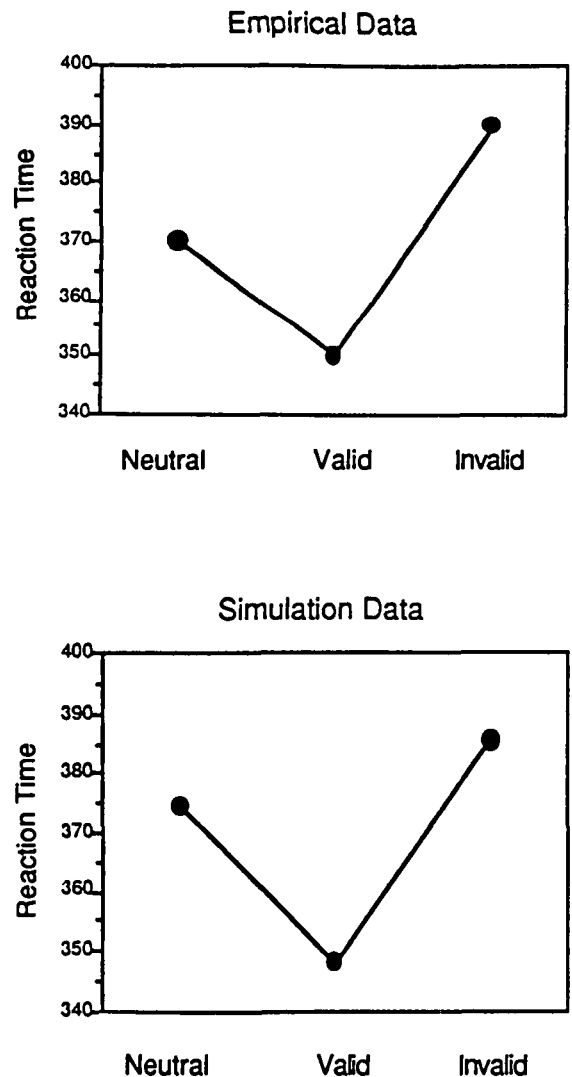


Figure 8. Performance of normal subjects and network in lateralized cued simple reaction time task. Number of cycles needed for “response” unit to reach threshold has been regressed onto reaction times.

cue and top-down activation that the attention units give the perception unit at its corresponding location. When the perception unit is activated by the cue, it activates the attention units on that side, which feed activation back to the perception unit, establishing a resonance that strengthens the activation of the target representation. Attentional inhibition due to invalid cuing is the result of the activated attention unit at the cued location suppressing the activation of the attention unit at the target location, leading to diminished top-down activation of the target perception unit. That is, the attention units on the cued side inhibit the attention units on the opposite side. As a result, when the target is presented to the opposite side, the attention unit on that side must first overcome the inhibition of the attention unit on the cued side before it can establish a resonance with its perception unit, and response time is therefore prolonged.

This very simple model of attention, which has no disengage component, captures the qualitative relations among the speeds of response in the three different conditions and can be fitted quantitatively to these average speeds with fairly good precision. In this regard, it seems preferable to a model that postulates separate components for orienting, engaging, and disengaging attention. The disengage component, however, was postulated on the basis of the behavior of parietally damaged subjects, not normal subjects. The critical test of this model, therefore, is whether it produces a disengage deficit when damaged.

A subset of the attention units on one side was eliminated and the model was run in the valid and invalid cuing conditions. (No patient data were available for the neutral condition.) Figure 9 shows the data of Posner et al. (1984) from parietally damaged patients and the simulation results, fitted to the data in the same way as before. Both sets of results show a disengage deficit: a disproportionate slowing from invalid cuing when the target is on the damaged side.

Why does the model show a disengage deficit when its attention units are damaged? The answer lies in the competitive nature of attentional allocation in the model and the imbalance introduced into the competition by unilateral damage. Attentional allocation is competitive, in that once the attention units on one side have been activated, they inhibit attentional activation on the other side. When there are fewer attention units available on the newly stimulated side, the competition is no longer balanced and much more bottom-up activation will be needed on the damaged side before the remaining attention units can overcome the inhibition from the attention units on the intact side to establish a resonance with the perception unit.

One might wonder whether we have really succeeded in simulating the disengage deficit without a disengage component, or whether some part of the model with a different label, such as the attention units or the inhibitory connections between attention units, is actually the disengage component. To answer this question, consider some of the attributes that would define a disengage component. First, it should be brought into play by perception of the target, and not the cue, on a given trial. Second, it should be used to disengage attention and not for any other function. By these criteria, there is no part of

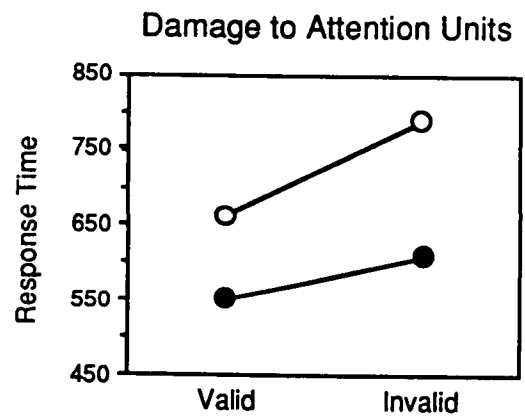
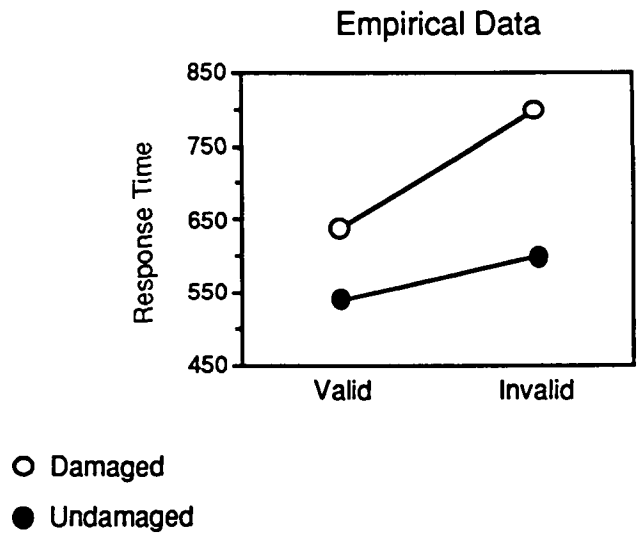


Figure 9. Performance of parietally damaged patients and damaged network in lateralized cued reaction time task. Number of cycles needed for "response" unit to reach threshold has been regressed onto reaction times.

the model that is a disengager. The attention units as well as their inhibitory connections are brought into play by both cue and target presentations. In addition, the attention units are used as much for engaging attention as for disengaging it. We therefore conclude that the disengage deficit is an emergent property of imbalanced competitive interactions among remaining parts of the system that do not contain a distinct component for disengaging attention.

Humphreys and Riddoch (1993) have independently proposed an account of the disengage deficit that does not include a disengage component in the normal architecture. Instead, they suggest that the deficit could be secondary to an impairment in orienting attention or to an overly strong engagement of attention ipsilaterally.

2.2.4. Relevance of the locality assumption for architecture of visual attention. After damage to the attention units on one side of the model, the nondamaged attention units on the other side function differently. Specifically, once activated they show a greater tendency to maintain their activation. This is because of the reduced ability of the attention units on the damaged side to recapture activation from the intact side, even when they are receiving bottom-up stimulus activation. The ability of this

model to account for the disengage deficit depends critically upon this nonlocal aspect of its response to damage.

2.3. The functional architecture of visual face recognition: Separate components for visual processing and awareness?

Prosopagnosia is an impairment of face recognition that can occur relatively independently of impairments in object recognition and is not caused by impairments in lower-level vision or memory. Prosopagnosic patients are impaired in tests of face recognition such as naming faces or classifying them according to semantic information (such as occupation); they are also impaired in everyday life situations that call for face recognition. Furthermore, based on their own introspective reports, prosopagnosics do not feel as though they recognize faces; however, when tested using certain indirect techniques, some of these patients do show evidence of face recognition. This has been taken to imply that their impairment lies not in face recognition per se, but in the transfer of the products of their face-recognition system to another system required for conscious awareness. This in turn implies that different components of the functional architecture of the mind are needed to produce perception and awareness of perception.

2.3.1. Evidence for dissociated recognition and awareness of recognition. Three representative types of evidence will be summarized here. The most widely documented form of "covert" face recognition occurs when prosopagnosics are taught to associate names with photographs of faces. For faces and names that were familiar to the subjects prior to their prosopagnosia, correct pairings are learned faster than incorrect ones (e.g., de Haan et al. 1987b). An example of this type of finding is shown in Table 4. It seems to imply that, at some level, the subject must have preserved knowledge of the faces' identities. The other two types of evidence come from reaction time tasks. One measures speed of visual analysis of faces, in which subjects must respond as quickly as possible to whether two photographs depict the same face or different faces. Normal subjects perform this task faster with familiar than unfamiliar faces. Surprisingly, as shown in Table 5, a prosopagnosic subject showed the same pattern, again implying that he was able to recognize them (de Haan et al. 1987b). The last task to be reviewed is a kind of semantic priming task. Subjects must classify printed names as actors or politicians as quickly as possible, while on some trials photographs of faces are presented in the background. Even though the faces are

Table 4. Performance on correct and incorrect face-name pairings in a face-name relearning task

Trial:	1	2	3	4	5	6	7	8
Correct pairings	2	1	1	2	1	2	0	3
Incorrect pairings	0	0	0	1	1	0	0	0
Trial:	9	10	11	12				
Correct pairings	2	3	2	2				
Incorrect pairings	1	1	0	0				

Table 5. Speed of visual matching for familiar and unfamiliar faces (in msec)

	Familiar	Unfamiliar
Prosopagnosic subject	2,795	3,297
Normal subjects	1,228	1,253

irrelevant to the task subjects must perform, they influence reaction times to the names. Specifically, normal subjects are slowed in classifying the names when the faces come from the other occupation category. As shown in Table 6, the prosopagnosic patient who was tested in this task showed the same pattern of results, implying that he was unconsciously recognizing the faces fully enough to derive occupation information from them (de Haan et al. 1987a; 1987b).

2.3.2. Interpretation of covert recognition relative to the functional architecture of visual recognition and conscious awareness. The dissociation between performance on explicit tests of face recognition and patients' self-reporting of their conscious experience of looking at faces, on the one hand, and performance on implicit tests of face recognition on the other, has suggested to many authors that face recognition and the ability to make conscious use of it depend on different components of the functional architecture. For example, de Haan et al. (1992) interpret covert recognition in terms of the components shown in Figure 10, in which separate components of the functional architecture subserve face recognition and conscious awareness thereof. According to their model, the face-specific visual and mnemonic processing of a face (carried out within the "face processing module") proceeds normally in covert recognition, but the results of this process cannot access the "conscious awareness system" because of a lesion at location number 1.

2.3.3. Accounting for dissociated covert and overt recognition with an interactive architecture. Randy O'Reilly, Shaun Vecera, and I (Farah et al. 1993) have modeled overt and covert recognition using the five-layer recurrent network shown in Figure 11, in which the same set of so-called face units subserves both overt and covert recognition. The face input units subserve the initial visual representation of faces, the "semantic" units represent the semantic knowledge of people that can be evoked either by the person's face or by the name, and the "name" units represent names. Hidden units were used to help the network learn the associations among patterns of activity in each of these three layers. These are located between the "face" and "semantic" units (called the "face"

Table 6. Priming of occupation judgments (in msec)

	Baseline	Unrelated	Related
Prosopagnosic subject	1,565	1,714	1,560
Normal subjects	821	875	815

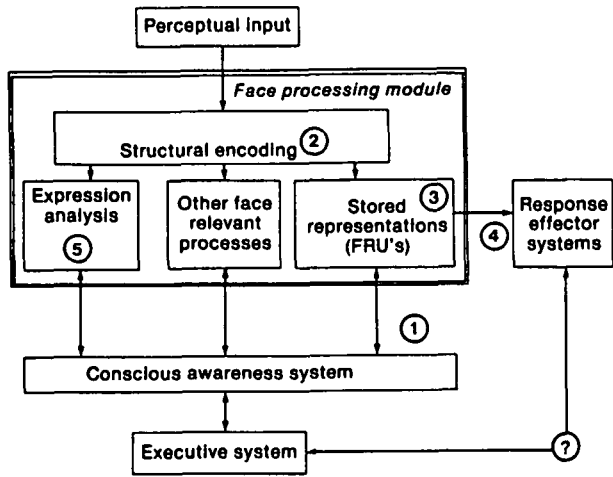


Figure 10. Functional architecture of perception and awareness, proposed by de Haan et al. (1992).

hidden units) and between the “name” and “semantic” units (the “name” hidden units). Thus, there are two pools of units that together comprise the visual face-recognition system in our model in that they represent visual information about faces: the “face input” units and the “face hidden” units.

The connectivity among the different pools of units was based on the assumption that in order to name a face, or to visualize a named person, one must access semantic knowledge of that person. Thus, face and name units are not directly connected but send activation to one another through hidden and semantic units. All connections shown in Figure 11 are bidirectional.

Faces and names are represented by random patterns of 5 active units out of the total of 16 in each pool. Semantic knowledge is represented by 6 active units out of the total of 18 in the semantic pool. The only units for which we have assigned an interpretation are the “occupation units” in the semantic pool: one represents the semantic feature “actor,” and the other, “politician.” The network was trained to associate an individual’s face, semantics, and name whenever one of these was presented, using the Contrastive Hebbian Learning algorithm (Movellan 1990). After training, the network was damaged by removing units.

Figure 12 shows the performance of the model in a 10-

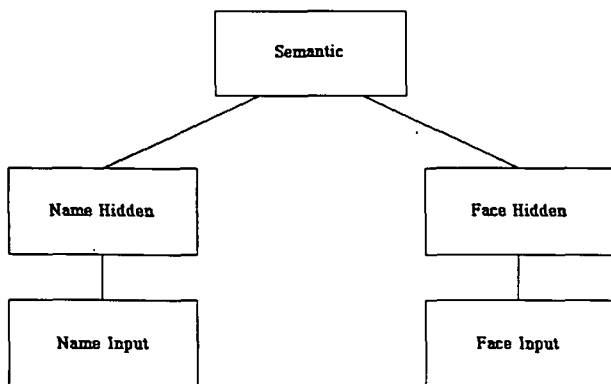
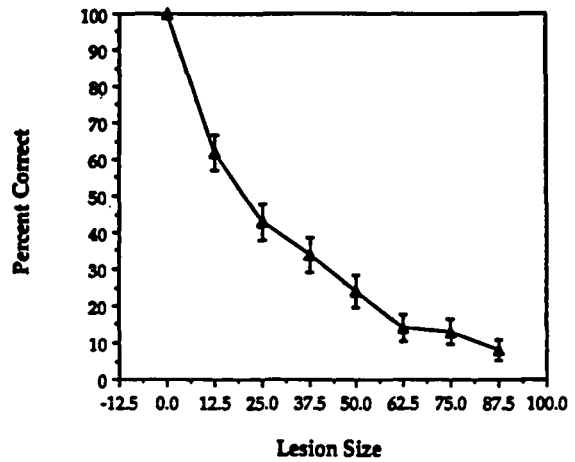


Figure 11. Functional architecture of face perception as modeled by Farah et al. (1993).

Overt Performance: Hidden Unit Lesions (Forced Choice with 10 Alternatives)



Overt Performance: Face Pool Lesions (Forced Choice with 10 Alternatives)

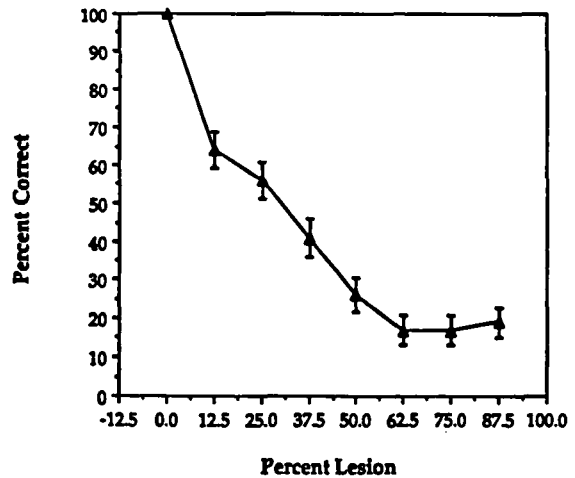


Figure 12. Effect of different amounts of damage to face units on the network’s ability to perform 10-alternative forced choice naming of faces, an overt face recognition task.

alternative, forced-choice naming task for face patterns after different degrees of damage to the “face input” and “face hidden” units. At levels of damage corresponding to removal of 62.5% and 75% of the face units in a given layer, the model performs at or near chance on this overt-recognition task. This is consistent with the performance of prosopagnosic patients who manifest covert recognition. Such patients perform poorly, but not invariably at chance, on overt tests of face recognition.

In contrast, the damaged network showed faster learning of correct face-name associations. When retrained after damage, it consistently showed more learning for correct pairings than incorrect ones in the first 10 training epochs, as shown in Figure 13. The damaged network also completed visual analysis of familiar faces faster than unfamiliar ones. When presented with face patterns after damage, the face units completed their analysis of the input (i.e., the face units settled) faster for familiar than unfamiliar faces, as shown in Figure 14. And finally, the damaged network showed semantic interference from faces in a name classification task. Figure 15 shows that when the network was presented with name patterns and

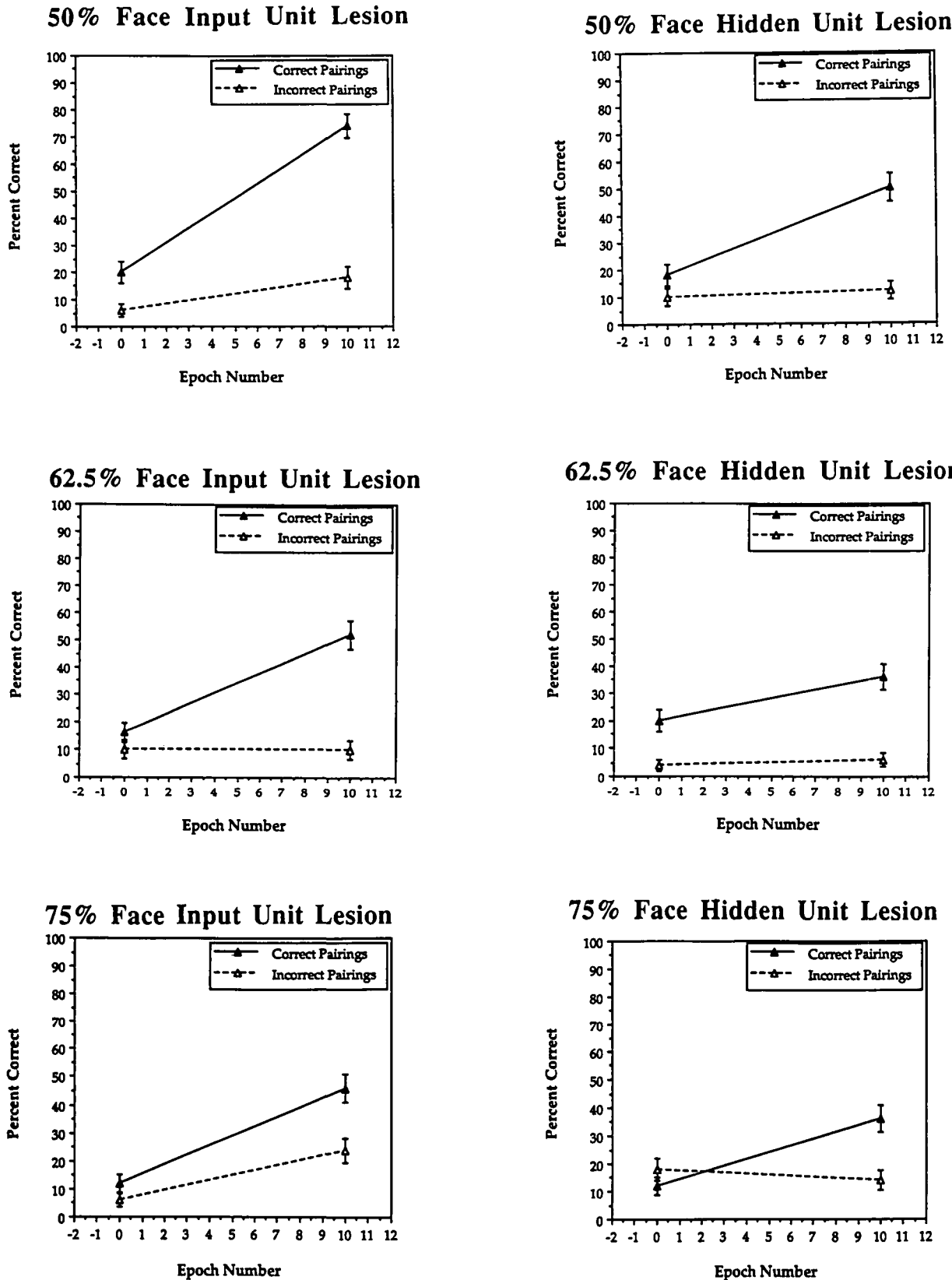


Figure 13. Ability of the network after different amounts of damage to face units to produce the name associated to a face (to within 2 bits), for correctly and incorrectly paired names and faces, immediately after damage and following 10 epochs of further training. Note that learning occurs more quickly for correctly paired names and faces.

the time it took to classify them according to occupation (i.e., the number of processing cycles for the occupation units to reach threshold) was measured, classification time was slowed when a face from the incorrect category

was shown, relative to faces from the correct category and, in some cases, to a no-face baseline.

Why does the network retain “covert recognition” of the faces at levels of damage that lead to poor or even

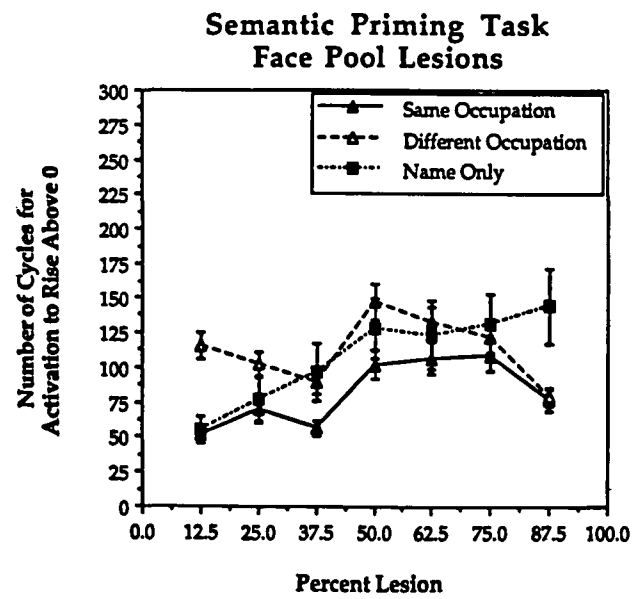
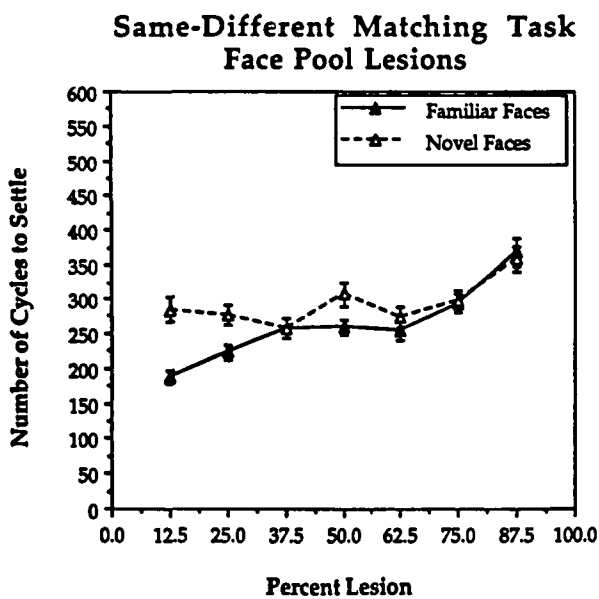
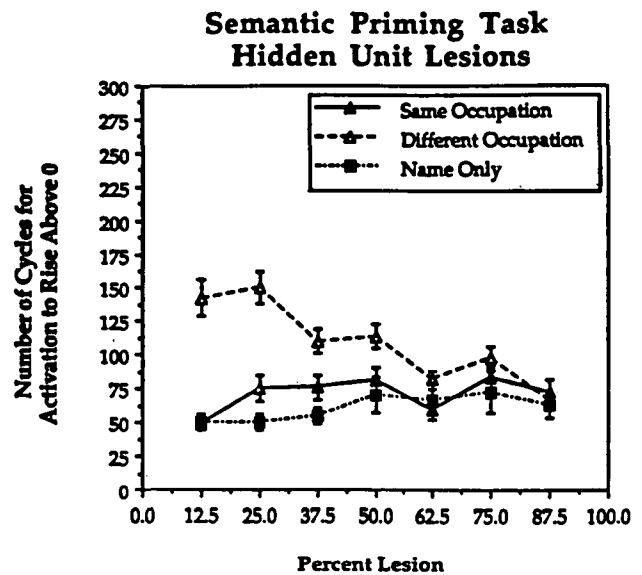
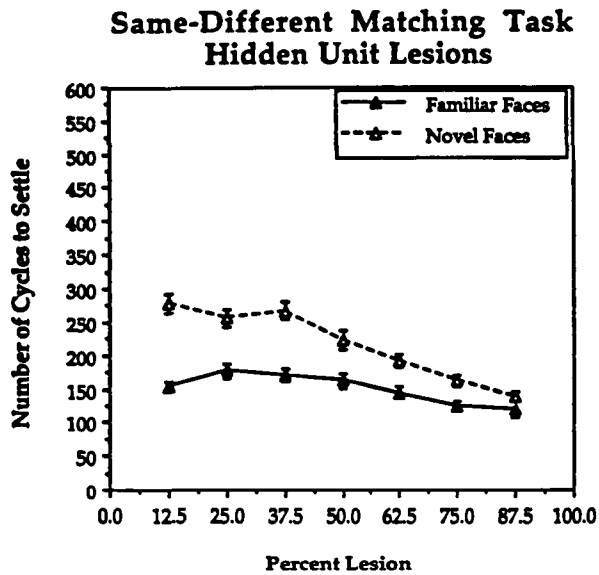


Figure 14. Effect of different amounts of damage to face units on the time needed for the face units to settle, for familiar input patterns (closed triangles) and for unfamiliar input patterns (open triangles). Note that familiar patterns tend to settle more quickly.

Figure 15. Effect of different amounts of damage to face units on the number of cycles needed for the "actor" and "politician" units to reach threshold when presented with name and face input patterns. When the face is from a different occupation category, it takes longer for the name to push the correct occupation unit over threshold.

chance levels of overt recognition? The general answer lies in the nature of knowledge representation in PDP networks. As already mentioned, knowledge is stored in the pattern of weights connecting units. The set of the weights in a network that cannot correctly associate patterns because it has never been trained (or has been trained on a different set of patterns) is different in an important way from the set of weights in a network that cannot correctly associate patterns because it has been trained on those patterns and then damaged. The first set of weights is random with respect to the associations in question, whereas the second is a subset of the necessary weights. Even if it is an inadequate subset for performing the overt association, it is not random; it has "embedded" in it some degree of knowledge of the associations. Furthermore, consideration of the tasks used

to measure covert recognition suggest that the covert measures should be sensitive to this embedded knowledge.

A damaged network would be expected to relearn associations that it originally knew faster than novel associations because of the nonrandom starting weights. The faster settling with previously learned inputs can be attributed to the fact that the residual weights come from a set designed to create a stable pattern from that input. Finally, to the extent that the weights continue to activate partial and subthreshold patterns over the nondamaged units in association with the input, these resultant patterns will contribute activation toward the appropriate units downstream, which are simultaneously being activated by intact name units.

2.3.4. Relevance of the locality assumption for architecture of perception and awareness. The role of the locality assumption is less direct in the foregoing example than in the previous two, but it is nevertheless relevant. Many authors have reasoned according to the locality assumption that the selective loss of overt recognition and the preservation of covert recognition implies that there has been localized damage to a distinct component of the functional architecture needed for overt, but not covert, recognition. The alternative account, proposed here, suggests that partial damage to the visual face-recognition component changes the relative ability of the remaining parts of the system (i.e., the remaining parts of the face-recognition component along with the other components) to perform the overt and covert tasks. Specifically, the discrepancy between the difficulty of the overt and covert tasks is increased, as can be seen by comparing the steep drop in overt performance as a function of damage shown in Figure 12 with the relatively gentle fall-off in the magnitude of the covert recognition effects shown in Figures 13–15. According to the model, this is because the information processing required by the covert tasks can make use of partial knowledge encoded in the weights of the damaged network and is therefore more robust to damage than the information processing required by the overt task. In other words, with respect to the relative ability of the remaining system to perform overt and covert tasks, the effects of damage were nonlocal. The ability of the model to account for the dissociation between overt and covert recognition depends critically on this violation of the locality assumption.

3. General discussion

3.1. Evaluating the truth and methodological necessity of the locality assumption

The foregoing examples were intended as a small “data base” with which to test two empirical claims about the locality assumption. First, that it is true, namely, that after local brain damage the remaining parts of the system continue to function as before. Second, that it is necessary, in other words, that there is no other way to make principled inferences from the behavior of brain-damaged subjects to the functional architecture of the mind, and that the only alternative is therefore to abandon cognitive neuropsychology.

The examples allow us to assess the likely truth of the locality assumption by assessing the likely truth of the different inferences made with and without it. Of course, each such pair of inferences was made on the basis of the same data and fits those data equally well, so the choice between them rests on considerations of parsimony and consistency with other information about brain organization. On the basis of these considerations, the inferences made without the locality assumption seem preferable. In the case of semantic memory, the model obtained without the locality assumption is consistent with an abundance of other data implicating modality-specificity as a fundamental principle of brain organization and with the lack of any other example of a purely semantic distinction determining brain organization. In the case of visual attention, the model obtained without the locality assumption has fewer

components: although the overviews of the models presented in Figures 6 and 7 are not strictly comparable (Fig. 6 includes components postulated to account for other attentional phenomena and Fig. 7 includes separate depictions of the left and right hemispheres’ attentional mechanisms as well as two different levels of stimulus representation), it can be seen that the same “attention” component shown in Figure 7 does the work of both the “engage” and “disengage” components in Figure 6. Similarly, setting aside the irrelevant differences in the complexity of Figures 10 and 11 arising from factors such as the greater range of phenomena to be explained by Figure 10, it is clear that the same visual “face” components in Figure 11 do the work of the visual “face” components and “conscious awareness system” in Figure 10, at least as far as explaining performance in overt and covert tasks is concerned.

It should be noted that the success of these models is a direct result of denying the locality assumption, as explained in subsections on the relevance of the locality assumption (sects. 2.1.4, 2.2.4, 2.3.4). In linking each neuropsychological dissociation to the more parsimonious functional architecture, a key explanatory role is played by the nonlocal effects that damage to one component of the architecture has on the functioning of other components. Hence the weight of evidence from the three cases discussed here suggests that the locality assumption is false. Finally, with respect to its necessity, the examples provide existence proofs that principled inferences can be made in cognitive neuropsychology without the locality assumption.

3.2. Possible objections

In this section I consider some possible objections to these conclusions, with the hope of clarifying what has and has not been demonstrated here.

3.2.1. PDP and box-and-arrow: Apples and oranges? One kind of objection concerns the comparability of the hypotheses that were derived with and without the locality assumption. The two types of hypotheses do indeed differ in some fundamental ways, and comparing them may be a bit like comparing apples and oranges. Nevertheless, apples and oranges do share some dimensions that afford meaningful comparisons, and I argue that the hypotheses under consideration here are likewise comparable in the ways discussed above.

For example, it might be objected that the computer models denying the locality assumption can only demonstrate the sufficiency of a theory, not its empirical truth, whereas the alternative hypotheses are empirically grounded. It is true that the models presented here have only been shown to be sufficient to account for the available data, but this is also true of the alternative hypotheses, and indeed of *any* hypothesis. It is always possible that a hypothesis can fit all the data collected so far, but that some other, as yet undiscovered, data could falsify it. The reason this may seem more problematic for PDP models is that there is a research tradition in computer modeling that takes as its primary goal the accomplishment of a task rather than the fitting of psychological data (e.g., Rosenberg & Sejnowski 1986), relying

exclusively on computational constraints rather than empirical constraints to inform the models. This is not a necessary feature of modeling, however, and the models presented here are constrained as much as the alternative hypotheses are by the empirical data.

Furthermore, the computational models presented here and the alternative hypotheses are on equal footing with respect to the distinction between prediction and retrodiction of data. In all three cases, the locality assumption has been used to derive a hypothesis, post hoc, from the observed neuropsychological dissociation. It was not the case that researchers had already formulated hypotheses to the effect that semantic memory was subdivided by taxonomic category or that there was a distinct component of the attention system for disengaging attention, or that awareness of face recognition depended on distinct parts of the mental architecture from face recognition; nor did they then go looking for the relevant dissociations to test those hypotheses. Rather, they began with the data and inferred their hypotheses just as we have done with the models presented earlier. Both the hypotheses derived using the locality assumption and the PDP models presented here await further testing with new data. An example of the way in which new data can be used to distinguish between the competing hypotheses comes from the work of Verfaellie et al. (1990) with a bilateral parietally damaged patient. They found that, contrary to their expectation of a bilateral disengage deficit, their subject showed diminished effects of attentional cuing. When attention units are removed bilaterally from the Cohen et al. (in press) model, which was developed before the authors knew of the Verfaellie et al. finding, the model also shows reduced attentional effects rather than a bilateral disengage deficit. This is because the disengage deficit in our model is caused by the imbalance in the number of attention units available to compete with one another after unilateral damage; bilateral damage does not lead to an imbalance but it does, of course, reduce the overall number of attention units and therefore the magnitude of the attentional effects.

Another way the comparisons presented above might seem mismatched is in their levels of description. The hypotheses derived using the locality assumption concern "macrostructure," that is, the level of description that identifies the components of the functional architecture, as shown in the so-called box-and-arrow models. In contrast, the hypotheses that deny the locality assumption appear to concern "microstructure," that is, the nature of the information processing that goes on within the architectural components. However, the latter hypotheses concern both microstructure and macrostructure, as should be clear from the macrostructures depicted in Figures 2, 7, and 11. We can therefore compare the two types of hypotheses at the level of macrostructure.

3.2.2. The locality assumption can be saved with more fine-grained empirical analysis of the deficit. Perhaps the prospects for the locality assumption look so dim because the types of data considered so far are unduly limited. The arguments and demonstrations presented above concern a relatively simple type of neuropsychological observation, namely, a selective deficit in some previously normal ability. I have focused on this type of observation for two reasons; the first is its very simplicity, and the seemingly

straightforward nature of the inferences that follow from it. At first glance, a truly selective deficit in *A* does seem to demand the existence of an *A* component, and this inference is indeed sound under the assumption that the *A* component is informationally encapsulated. The second reason is that this is still the most common form of inference in cognitive neuropsychology, as argued earlier in the section on ubiquity (sect. 1.2).

Nevertheless, other, finer-grained ways of analyzing patient performance are used increasingly by cognitive neuropsychologists to pinpoint the underlying locus of impairment in a patient's functional architecture. The two most common are qualitative error analyses, and selective experimental manipulations of difficulty of particular processing stages. Can the use of the locality assumption be buttressed by the additional constraints offered by these methods? Several recent PDP simulations of patient performance suggest that these finer-grained analyses are just as vulnerable to nonlocal effects of brain damage as are the more brute-force observations of deficit *per se*.

For example, semantic errors in single-word reading (e.g., pear → "apple") have been considered diagnostic of an underlying impairment in the semantic representations used in reading, and visual errors (pear → "peer") are generally taken to imply a visual processing impairment (e.g., Coltheart 1985). Hinton and Shallice (1991) showed how a PDP simulation of reading could produce both kinds of errors when lesioned either in the visual or the semantic components of the model. Humphrey et al. (1992) make a similar point in the domain of visual search: error patterns suggestive of an impairment in gestalt-like grouping processes can arise either from direct damage to the parts of the system that accomplish grouping or by adding noise to earlier parts of the system. In both cases, the nondiagnosticity of error types results from the interactivity among the different components of the model.

Another well-known example of the use of error types to infer the locus of impairment is the occurrence of regularization errors in the reading performance of surface dyslexics (e.g., Coltheart 1985). As mentioned earlier, surface dyslexics fail to read irregular words; this has been interpreted, using the locality assumption, as the loss of a whole-word reading route with preservation of the sublexical grapheme-phoneme translation route. The inference that these patients are relying on the latter route seems buttressed by a further analysis of the nature of their errors, which are typically regularizations (e.g., *pint* is pronounced like "lint"). Patterson et al. (1989), however, showed that a single-route architecture, comprised only of whole-word spelling-sound correspondences, produced, when partially damaged, both a selective impairment in the reading of irregular words and a tendency to regularize them. With the distributed representations used in their model, similar orthographies and phonologies have similar representations at each of these levels and there is consequently a tendency toward generalization. Although with training the system learns not to generalize the pronunciation of, say, *pint* to the pronunciation of most other *-int* words (such as *lint*, *mint*, *hint*), this tendency is unmasked at moderate levels of damage. The model's regularization errors are probably best understood as a result of the distributed nature of the word

representations in their model. The principles of PDP are closely interrelated, however, and the regularization effects can also be viewed as the result of interactions among different word representations, with the less common pronunciations losing their "critical mass" and therefore being swamped by the remaining representations of more common pronunciations.

Analyses of selective deficits and of the nature of the errors produced have in common the use of a purely observational method. Perhaps experimental manipulations designed to tax the operation of specific components offer a more powerful way of pinpointing the locus of impairment. Two recent models speak to this possibility and show that direct manipulations of particular processing stages are no more immune to nonlocal effects than are the previous methods. Mozer and Behrmann's (1990) model of visual-spatial neglect shows how the manipulation of a stimulus property designed to affect postvisual processing, namely the lexicality of a letter string (word, pseudoword, nonword), can have pronounced effects on the performance of a model whose locus of damage is visual. Interactions between attended visual information and stored lexical representations allow letter strings to be reconstructed more efficiently the more they resemble familiar words. Tippett and Farah (in press) showed how apparently conflicting results in the literature on the determinants of naming difficulty in Alzheimer's disease can be accounted for with a single hypothesis. Although most researchers believe that the naming impairment in Alzheimer's disease results from an underlying impairment of semantic knowledge, manipulations of visual difficulty (degraded visual stimuli) and lexical access difficulty (word frequency) have pronounced effects on patients' likelihood of naming, leading to alternative hypotheses of visual agnosia or anomia (Nebes 1989). When semantic representations were damaged, a PDP model of visual naming showed heightened sensitivity to visual degradation and word frequency. Thus, when one component of an interactive system is damaged, the system as a whole becomes more sensitive to manipulations of the difficulty of any of its components.

In sum, the problem of nonlocal effects of brain damage is not limited to inferences based on the range and boundaries of the impairment; it also affects inferences based on the qualitative mode of failure and the sensitivity of the system to manipulations designed to affect specific components directly.

3.2.3. PDP could be false. A different type of objection concerns the assumptions of the PDP framework. As already acknowledged, PDP is controversial. How can one be convinced, through comparisons involving PDP models, that the locality assumption is false, if it has not been established first that PDP is a correct way of characterizing human information processing? First, it should be pointed out that much of the controversy concerning PDP involves the adequacy of PDP models of language and reasoning, which are not relevant here. Few vision researchers would deny that the basic principles of PDP are likely to apply to visual attention and pattern recognition (e.g., see the recent textbook overviews of these topics by Allport 1989; Biederman 1990; Hildreth & Ullman 1989; Humphreys & Bruce 1989; and even Pinker 1985, who has been critical of PDP models of language).

Semantic memory may be a more controversial case. Second, and perhaps more important, one can remain agnostic about PDP as a general framework for human information processing and still appreciate that the particular models presented here are credible alternatives to those derived using the locality assumption. PDP, like the locality assumption, is ultimately an empirical claim that will gain or lose support according to how well it helps explain psychological data. The ability of PDP to provide parsimonious accounts for neuropsychological dissociations such as the ones described here counts in its favor. Finally, even if PDP were false, there would remain other ways of conceptualizing human information processing that would provide explicit, mechanistic alternatives to modularity. For example, in production system architectures (see Klahr et al. 1987) working memory is highly nonencapsulated. Kimberg and Farah (in press) found that weakening association strengths in working memory produced an array of specific and characteristic frontal impairments that were in no transparent way related to working memory. Although interactive computation is at the heart of PDP, which makes PDP the natural architecture to contrast with the locality assumption, other architectures are also capable of accommodating high degrees of interactivity.

3.3. General implications of denying the locality assumption

3.3.1. Modularity. The truth of the locality assumption has implications for issues in psychology beyond how best to infer functional architecture from the behavior of brain-damaged patients. As discussed at the outset, the locality assumption follows from a view of the mind and brain according to which the components of the functional architecture are informationally encapsulated, that is, their inputs and outputs are highly constrained. Components interact only when one has completed its processing, at which point it makes the end product available to a relatively small number of other components. If this were true, then the effects of damaging one component should be relatively local. Alternatively, if we judge that the best interpretation of various neuropsychological deficits (on the grounds of parsimony or consistency with other scientific knowledge, not on the grounds of a priori preferences for encapsulation or interactivity) involves denying the locality assumption, then this counts as evidence against modularity.

The term "modularity" is often used in a more general sense than I have used it so far, and this more general sense is not challenged by the failure of the locality assumption. Specialized representations are sometimes called "modules," so that the model in Figure 2 could be said to contain "visual knowledge" and "functional knowledge" modules. In this more general sense, the "modularity hypothesis" is simply that there is considerable division of labor among different parts of functional architecture with, for example, knowledge of language represented by a separate part of the system (functionally, and possibly anatomically), compared with other knowledge. Of course, if such a system is highly interactive, it may be difficult to delineate and characterize the different modules, but this is a problem of *how* you find something out, not of what it is or whether it exists.

3.3.2. Top-down versus bottom-up research strategies.

Denying the locality assumption also has a more general implication for research strategy in cognitive neuroscience. Most researchers in neuroscience and cognitive science acknowledge that there are multiple levels of description of the nervous system, from molecules to thoughts, and that one of the goals of science is a complete description of the nervous system at all of these levels. However, such researchers may differ in their opinions as to the most efficient way to arrive at this complete description. The bottom-up, or reductionist, approach is to begin with the most elementary levels of description, such as the biophysics of neurons, believing that it will be impossible to understand higher levels of organization if one does not know precisely *what* is being organized. This approach is anathema to cognitive neuroscience, which is, by definition, forging ahead with the effort to understand such higher-level properties of the brain as perception, memory, and so forth, while acknowledging that our understanding of the more elementary level of description is far from complete.

The main alternative, explicitly endorsed by many cognitive neuroscientists, is the top-down approach, according to which the most efficient way to understand the nervous system is by successive stages of analysis of systems at higher levels of description in terms of lower levels of description. It is argued that our understanding of lower levels will be facilitated if we know what higher-level function they serve. It is also argued that the complexity of the task of understanding the brain will be reduced by the "divide and conquer" aspect of this strategy, in which the system is analyzed into simpler components that can then be further analyzed individually (e.g., Kosslyn et al.'s 1990 "hierarchical decomposition constraint"). In the context of the three examples discussed earlier, this corresponds to first deriving the macrostructural hypotheses, in which the relevant components of the functional architecture are identified, and then investigating the microstructure of each component's internal operation. Unfortunately, to derive a macrostructure from neuropsychological data requires either making the locality assumption or considering the system's microstructure, as was done in the foregoing examples. If the locality assumption is false, the microstructure has implications for the macrostructure, and one cannot be assured of arriving at the correct macrostructural description without also considering hypotheses about microstructure.

Thus, even if one's only goal is to arrive at the correct macrostructural description of the functional architecture, as is the case for most cognitive neuropsychologists, the three examples presented here suggest that one must nevertheless consider hypotheses about microstructure. This points out a correspondence between theories of functional architecture and the methodologies for studying it. If one holds that the components of the functional architecture are informationally encapsulated, one can take a strictly top-down approach to the different levels of description, "encapsulating" one's investigations of the macrostructure from considerations of microstructure. In contrast, if one views the functional architecture as a highly interactive system, with each component responding directly or indirectly to the influences of many others, then one must adopt a more interactive mode of research,

in which hypotheses about macrostructure are influenced by constraints imposed simultaneously at both the macrostructural and the microstructural levels.

3.3.3. Implications for cognitive neuropsychology.

The conclusion that the locality assumption may be false is a disheartening one. It undercuts much of the special appeal of neuropsychological dissociations as evidence about the functional architecture. Although perhaps naive in hindsight, this special appeal came from the apparent directness of neuropsychological data. Conventional methods of cognitive psychology are limited to what Anderson (1978) has called "input-output" data: manipulation of stimuli and instructions on the input end and the measurement of responses and response latencies at output. From the relations between these, the nature of the intervening processing must be inferred. Such inferences are indirect, and as a result often underdetermine choices between competing hypotheses. In contrast, brain damage directly affects the intervening processing, constituting a direct manipulation of the "black box."

Unfortunately, the examples presented here suggest that even if the manipulation of the intervening processing is direct, the inferences by which the effects of the manipulations must be interpreted are not. In Ferrier's (1886) words, it may well be "at least highly difficult to trace any uncomplicated connection between the symptoms produced and the lesion as such." The locality assumption, which constitutes the most straightforward way of interpreting neuropsychological impairments, does not necessarily lead to the correct interpretation. If the locality assumption is indeed false, then dissociations lose their special status as particularly direct forms of evidence about the functional architecture.

Even for cognitive neuropsychologists who would not claim any special status for neuropsychological data, abandoning the locality assumption would make their work harder. The interpretation of dissociations without the locality assumption requires exploring a range of possible models that, when damaged, might be capable of producing that dissociation. What makes this difficult is that the relevant models would not necessarily have components corresponding to the distinctions between preserved and impaired abilities and we therefore lack clear heuristics for selecting models to test.

The foregoing demonstrations and arguments are not intended to settle decisively the issue of whether the locality assumption is correct. As already acknowledged, this is not the type of issue that can be decided on the basis of a single study or even a small number of studies. Instead, my goal has been to call attention to the fact that we do not have any firm basis for an opinion one way or the other, despite the widespread use of the locality assumption. Furthermore, at least in a few cases the best current interpretation seems to involve denying the locality assumption.

It is possible that some cognitive domains will conform more closely to the locality assumption than others; if so, this would have interesting theoretical as well as methodological implications concerning the degree of informational encapsulation in different subsystems of the functional architecture. However, until we have a broad enough empirical basis for deciding when the locality assumption can safely be used and when it will lead to

incorrect inferences, we cannot simply assume it to be true, as has been done almost universally in the past.

ACKNOWLEDGMENTS

The writing of this paper was supported by ONR grant N00014-91-J1546, NIMH grant R01 MH48274, NIH career development award K04-NS01405, and a grant from the McDonnell-Pew Program in Cognitive Neuroscience. I thank my coauthors on the projects described herein for their collaboration and tutelage in PDP modeling: Jonathan Cohen, Jay McClelland, Randy O'Reilly, Rick Romero, and Shaun Vecera. Special thanks to Jay McClelland for his encouragement and support. I thank several colleagues for discussions of the ideas in this paper: John Bruer, Alfonso Caramazza, Clark Glymour, Mike McCloskey, Morris Moscovitch, Edmund Rolls, and Larry Squire. I also thank Larry Weiskrantz for calling my attention to the passage from Ferrier quoted at the beginning. Finally, I thank the reviewers and editor of *BBS* for useful comments and criticisms of a previous draft: C. Allen, S. Harnad, G. Humphreys, M. McCloskey, M. Oaksford, T. Van Gelder, and four anonymous reviewers.

NOTES

1. There are, of course, many other ways to make a wrong inference using the locality assumption, even with the foregoing conditions satisfied, but these have to do with the particular content of the hypothesis being inferred and its relation to the data, not the use of the locality assumption per se. For example, Caramazza et al. (1990) have pointed out that selective impairments in modality-specific knowledge do not imply that knowledge of different modalities is represented in different formats; dissociability will not, in general, tell us about representational formats.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Simulating nonlocal systems: Rules of the game

John A. Bullinaria

Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, Scotland

Electronic mail: johnbull@ed.ac.uk

Farah notes that the locality assumption, as usually stated, may appear "naive." Indeed, Shallice (1988, Ch. 11; see also multiple review, *BBS* 14[3] 1991) has described a whole range of systems that can give rise to double dissociations (DDs), and Dunn and Kirsner (1988) have presented a class of single process systems that can result in DDs and have formulated the new concept of reversed association to replace DDs as a valid indicator of modularity. Consequently, one might argue that the locality assumption is simply an artefact from the days of "box-and-arrow" models and that since the advent of connectionist (i.e., parallel distributed processing [PDP] or neural network) modelling we can (and should) work in finer detail and use finer-grained evidence to constrain our theories.

As Farah notes, for a fair comparison of local and nonlocal (or,

box-and-arrow and connectionist) systems we must place appropriate constraints on both approaches. To make general statements about the need for the locality assumption and the superiority of connectionism over boxes and arrows we should also constrain the types of system we attempt to model. Thus, in addition to the implicit constraint that our nonlocal systems are neural networks, it seems reasonable to propose the following "rules of the game":

1. Since single dissociations and weak DDs can be explained as resource artefacts (Shallice 1988, Ch. 10), and no one will be impressed if we use anything more complicated, we should concentrate on modelling strong DDs.

2. If a neural network is simple enough to set the connection weights (synaptic strengths) by hand then a PDP approach is probably not required. The weights should be learnt.

3. If the system has inbuilt structure it is little more than an explicit implementation of a box-and-arrow system. We should try to allow any modularity to be learnt rather than imposed by hand. However, if we then fail to find DDs we will not know whether this is because DDs in real brains arise solely because of innate structures that have not been built into our models or if our learning algorithms are too dissimilar to those in real brains for the same modular structures to arise.

4. If we are forced to have inbuilt structure then with neural networks we no longer have to restrict ourselves to cases with truly separate modules – we can (and should) try all kinds of subtle connections between the modules, examining their implications.

5. We must ensure that our neural networks are sufficiently complex: we need enough training patterns to prevent them from operating by table lookup and enough hidden layers and connections for a modular structure to arise if that is appropriate.

6. The input and output representations must be chosen carefully – in small systems random fluctuations can be converted into strong dissociations by an (un)suitable choice of representation.

7. We often restrict our neural networks to the minimum number of units and connections required to solve the problem because this tends to speed up the training, improves generalisation, and makes it easier to understand the hidden unit representations. Minimal systems are not likely to behave in the same way as nonminimal systems such as brains, however, and should be avoided.

8. Some forms of neural network damage are more realistic than others. The most obvious is the removal of subsets of units and connections but we should also consider various changes to the weights and activations: adding noise, global rescaling, clipping, and so on.

9. Neurological patients often (but not always) show rapid improvement in performance after a lesion occurs (Geschwind 1985) so we should also allow our systems to relearn after damage. With minimal networks we can easily lesion them so that they become subminimal and then allow them to relearn. This can confuse the results since relearning can create, destroy, or even reverse the sense of dissociations. For nonminimal networks we have the problem that relearning tends to compensate totally for the damage and we get no dissociations at all.

10. One must be able to argue that the system and our analyses of it can scale up to sizes comparable to those found in the brain. Combinatorial explosions often make this difficult. Many of these points are discussed in more detail in Bullinaria and Chater (1993). These rules are difficult to follow and should perhaps be aimed for in the future rather than expected of current models. None of Farah's three models gets past rule 1 and it is unlikely that any existing model satisfies them all.

We end by suggesting the kind of model that might one day satisfy these rules. We deal explicitly with reading (where surface and phonological dyslexia constitute a strong DD) since this was listed by Farah as a typical case where locality is

assumed. In fact, there have been arguments against locality here (in particular the strong dual-route hypothesis) for some time, even before PDP came along (e.g., Humphreys & Evett 1985). There now exist fully distributed, nonminimal, single-route neural network models of reading that learn to achieve 100% performance on both regular and exception words of any length in their training data and can read nonwords as well as humans (Bullinaria 1993). With a specific form of global damage (that easily scales up) they exhibit symptoms similar to surface dyslexia but are unable to acquire phonological dyslexia, suggesting that there must still be some kind of missing (lexical/semantic) route. These models can already use context information to deal successfully with homographs so, although it remains to be seen how we can fully incorporate semantics into these systems, any additional route is unlikely to be totally independent and uncoupled. We are within the rules so far. If we can add in the semantics and get the full DD without breaking the rules we might really have a serious and general challenge to the locality assumption.

Local representations without the locality assumption

A. Mike Burton and Vicki Bruce

Department of Psychology, University of Stirling, Stirling FK9 4LA, United Kingdom

Electronic mail: mb1@forth.stir.ac.uk; vb1@forth.stir.ac.uk

Farah's arguments against what she calls the "locality assumption" appear to us to be well-founded, yet we would like to take issue with two aspects of her target article. First, we will argue that the rejection of this assumption should not lead one to take on board all the assumptions of the parallel distributed processing (PDP) approach. We focus in particular on the "distributed" assumption. Second, we will argue that many cognitive neuropsychologists share the same insight about the locality assumption. As a result, much recent theoretical work avoids this assumption but does not necessarily resort to a radical PDP approach.

PDP as a "specific" alternative? Farah lists three assumptions of the PDP approach: distributed representations, gradedness, and interactivity (sect. 1.4). She reminds us that the psychological plausibility of PDP is controversial, stating that "PDP is to be evaluated as a *specific* alternative [to the locality] assumption" (sect. 1.4, our italics). The three example models used to test the PDP approach, however, seem radically different from one another. It appears that PDP is not a *specific* alternative, but a huge family of alternatives that may be mixed and matched as it suits the builders of these models. It is never clear how much of the particular PDP architecture of each model is intended to carry explanatory power in accounting for the phenomenon under study. We will focus on the issue of distributed representations to illustrate this point.

In the model of semantic memory (an autoassociator trained with the delta rule), the notion of a distributed representation is used in a componential way, so a representation is highly visual if it comprises a higher proportion of visual semantic units than functional semantic ones. Under some (strict) definitions of distributed representations, individual units have no referent. In this case, however, each of the semantic units must in some sense be individually referential, as their effect can be summed. This componential approach is not new. If by the term distributed Farah means that large things are made up of smaller things, we are not prepared to dispute this. However, this is not the exclusive insight of the PDP approach.

In the model of visual attention (a snapshot model with no learning) there is no very clear definition of whether representations are distributed or local, but the model of face perception

(trained with the contrastive Hebbian learning algorithm) clearly mixes local and distributed representations. The input to the model from either face or name is a distributed pattern over 5 (out of 16) input units. These input patterns appear to be more strictly distributed than in the model of semantic memory. As far as we can see, no content is assigned to any of the individual components of these input patterns. The semantic units, on the other hand, appear to be entirely local, as two of these are assigned the interpretation "actor" and "politician."

These examples illustrate a weakness in the PDP approach as articulated by Farah. The assertion that explanations of cognitive phenomena should be distributed is insufficiently constraining. There are infinitely many ways we may construct distributed representations, as opposed to the single way we construct local representations. Unless advocates of PDP modelling are prepared to state exactly *how* representations should be distributed, there is little to be gained from this dictum. Researchers will inevitably choose a mode of distribution which allows their model to work, but the reasons for this choice will usually remain hidden. Appeals to the undeniably distributed nature of neural processing are irrelevant here, unless researchers are explicitly modelling particular structures in the brain.

Finally, there are pragmatic reasons to use local representations where possible. Distributed representations are inherently difficult to interpret. There are various ways one can interrogate a model comprising this type of representation (e.g., dot-products with canonical representations, solution of simultaneous equations, etc.). However, one cannot simply observe their behaviour directly. This of course adds to a model's mystique, but mystique is not necessarily desirable in cognitive theorising.

Does the locality assumption prevail? Contrary to Farah's statement at the start of section 1.1, cognitive neuropsychologists do not "generally assume" that damage to one component of the functional architecture will have exclusively local effects. Many cognitive neuropsychologists are exploring the potential of PDP models as ways of implementing theories of cognitive processing; Farah's text is peppered with examples of this (e.g., Hinton & Shallice 1991; Humphreys et al. 1992; Patterson et al. 1989).

In (at least) the case of covert recognition in prosopagnosia, Farah seems to have focussed selectively on an example of "old-fashioned" cognitive neuropsychologising. It is of course possible to offer an account of the phenomenon without resorting to an explanation in terms of a separate "awareness" component. Such an account in fact already exists in the literature (Burton et al. 1991). This model uses an interactive activation and competition (IAC) architecture (cf. McClelland 1981), which includes graded responses and interactivity (but not distributed representations). Moreover, that account is not based on "a hypothesis [derived] post hoc from the observed neuropsychological dissociation" (cf. Farah, sect. 3.2.1). The account of covert recognition was built upon a preexisting theoretical framework for face processing and person identification (Bruce & Young 1986, itself a revision of earlier work by Hay & Young 1982). This framework has been built on converging evidence from experimental psychology and neuropsychology, and in its latest stage of development has implemented part of the framework in IAC terms (Bruce et al. 1992; Burton et al. 1990; 1991). We did not need to overturn the apple (orange?) cart to develop interactive simulations; instead, these form a natural progression from the "box-and-arrow" style of theorising which preceded them. One of our worries about the *style* of PDP theorising exemplified in Farah's target article is that it seems to represent *punctate* modelling of isolated phenomena. We find models (of any architecture) more constructive when they build upon and accommodate previous bodies of data and theory derived from diverse sources.

In conclusion, we agree with Farah's call for gradual and interactive models in neuropsychological and cognitive theorising. Many cognitive neuropsychologists appear to share this

view. However, we take issue with the assertion that these models need to be distributed (except in the trivial sense of compositionality). Local representations in such simple architectures as IAC networks carry all the advantages of the interactive approach, but none of the problems associated with distributed representations.

Regional specialities

Brian Butterworth

Department of Psychology, University College London, London WC1E 6BT, United Kingdom

Electronic mail: ucjtsbb@ucl.ac.uk

The first puzzling feature of Farah's "critique" is this: Who is being criticised? She claims that the locality assumption (LA) is ubiquitous among neuropsychologists. She cites four researchers in four domains who are meant to exemplify this commitment, but she does not demonstrate that they "assume that damage to one component of the functional architecture will have exclusively 'local' effects"; nor that any of the researchers assume that these components are informationally encapsulated. I can think of no one who has maintained in print that all components of the brain's functional architecture are informationally encapsulated. Even Fodor (1983) separates modularised – and hence encapsulated – input systems from nonencapsulated central systems. This is a distinction Farah does not mention. Why is information encapsulation picked on as the sole defining characteristic of the neuropsychologists' functional components? Why not domain-specificity, which would be a much more plausible choice for most neuropsychologists?

When she comes to consider neuropsychologists' explicit accounts of their own methodologies, Farah details the work of two influential methodological theorists who, by her own admission, *do not* hold the LA: Shallice's (1988) account of "isolable subsystems" entails, according to her, that "the locality assumption is not strictly true"; and Caramazza's (1986) "transparency principle" is "probably weaker than the LA." They are nevertheless attacked as if they do hold LA.

Neuropsychologists – not to mention neurologists and neuroscientists – certainly believe that regions of the brain may have specialised functions. They do not *assume* this in order to make neuropsychology possible: they infer it from the abundant phenomena. The second puzzling feature, therefore, is this: Does Farah believe that regions of the brain do not have specialisations? Does she believe in the theory of "mass action"? She writes sometimes as if she did: her "constraining principles" of distributed representation, graded information processing, and interactivity are consistent with this theory.

A third puzzling feature: Farah is inviting us to reject the LA, if we ever held it, in favour of a parallel distributed processing (PDP) approach, on the grounds that theories based on three PDP simulations do a better job explaining neuropsychological data than previous non-PDP accounts. Now, of course, the fact that *some* A (normal performances) are B (best explained by PDP), does not entail that *all* A are B. Even to licence the inference from one better PDP account to a rejection of the universal necessity of the LA she needs to show that the inferior explanation assumes locality in Farah's sense.

Certainly this is not the case for Warrington and Shallice's story about category-specific deficits (Shallice 1988, Ch. 12; Warrington 1975, 1981). Warrington and McCarthy (1987) quite explicitly entertain differential weightings for different types of information. This, they argue, is necessary to explain fine-grained within-category effects – such as selective deficits for fruit and vegetables (e.g., Hart et al. 1985). Thus

colour, shape, motion and location are known to be separable both physiologically, anatomically and psychologically at very early stages

of information processing. . . . It seems not implausible to suggest that this early segregation may have concomitants for later stages of cognitive analysis and that the evidence provided by or derived from such functions, i.e. channels, may *interact differentially with information from other sources*. (p. 1291; my italics)

The processing of information about living things is, for them, though perhaps localised, explicitly *not* informationally encapsulated. Shallice (1988, p. 302) notes, regarding category specificity: "Instead of conceiving of the semantic system as a set of discrete subsystems (functions, sensory properties, and so on) . . . it may be more useful to think of it as a giant distributed net in which regions tend to be specialised for different types of processes." So an attack on standard neuropsychological accounts of category-specific deficits would not be an attack on a position here that entails the LA.

If Farah wishes to persuade neuropsychologists of the usefulness of PDP modelling, she needs to demonstrate that a single network can be selectively lesioned to produce their favourite type of observation: the double dissociation. Has she done this? Her treatments of attention and face recognition deal only with a single dissociation. Her only case of modelling a double dissociation is the category-specific losses in semantic memory, where she shows selective impairment of living (Fig. 3, top panel) and nonliving things (Fig. 3). This has been achieved by selectively damaging one set of units ("visual semantic memory units") to get a deficit of living things, and another set of units ("functional semantic memory units") to get the deficit of nonliving things. This is not formally different from the position she is attacking. The architecture of the model in Figure 1 is identical to the one in Figure 2. Only the labels have been changed. It may well be, as she has argued in Farah and McClelland (1991; and indeed as Warrington and McCarthy had argued on the basis of patient data), that labelling the nodes "functional" and "visual" is to be preferred, but this cannot bear on the question of architecture or theoretical framework. There are still sets of nodes dedicated to specific functions. Like the rest of us neuropsychologists, she explains double dissociations in terms of selective damage to two separable systems – in neural terms, two regions, each with its own speciality.

Locality, modularity and numerical cognition

Jamie I. D. Campbell

Department of Psychology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0

Electronic mail: campbellj@sask.usask.ca

Farah does not discuss numerical cognition in her target article, but much recent research on number processing in normal and brain-damaged subjects has focused on the issue of locality, modularity, and selective impairments. In this commentary I outline how the locality assumption has been applied to numerical cognition and describe phenomena from normal subjects that, consonant with Farah's conclusions, challenge the locality assumption and the simple type of dissociation logic it promotes.

Architectures for cognitive number processing

Modular model. The locality assumption is inherent in one of the prominent theories of cognitive number processing and dyscalculia, the modular model introduced by McCloskey et al. (1985; see McCloskey 1992 for a review of related research). According to the model, basic number processing is comprised of three functionally distinct groups of processes specialized for number comprehension, calculation, and production, respectively. The comprehension system converts different surface notations (e.g., digits, written, or spoken number words) into an abstract (i.e., modality-independent) semantic code that provides the basis for subsequent processing in the calculation or

production systems. The calculation system stores numerical information in the abstract format and includes memory for basic numerical facts, rules, and procedures (e.g., $6 + 7 = 13$, $6 \times 9 = 54$, $0 \times N = 0$, $0 + N = N$). The production system is comprised of modules that convert the abstract output from the comprehension or calculation modules into specific output formats, such as digits or written or spoken number words.

The locality assumption is inherent in the modular model because it is assumed that (1) the hypothetical modules are represented separately and so can be disrupted independently by brain damage, and (2) no component process is involved in the inner workings of the others (McCloskey et al. 1992, pp. 496–97). McCloskey (1992) reviews case studies that identify selective deficits corresponding to the major subsystems in the modular model and that, given the locality assumption, appear to validate the functional modular architecture assumed in the model. For example, McCloskey, et al. (1986) observed patients with impaired number-reading skills who could successfully perform certain numeral transcoding, comparison, and calculation tasks. Conversely, other patients were observed with number-reading and number-comparison skills relatively intact, but with impaired calculation (Sokol et al. 1991). Such dissociations seem to support the view that calculation processes are independent of the encoding and production mechanisms that mediate number reading.

Interactive model. Campbell and Clark (1988; 1992; Campbell 1992; 1993; Clark & Campbell 1991) challenged the locality assumption on which the modular model is based. They proposed an alternative *encoding complex* theory in which numerical encoding and calculation functions are integrated and interactive and depend on modality-specific processes rather than abstract codes. One compelling source of evidence for the encoding-complex view came from analyses of normal adults' speed-induced errors of simple addition and multiplication (Campbell 1992; 1993; Campbell & Clark 1992). The errors frequently involved *operand intrusions* in which a problem operand appeared in the error response (e.g., $2 + 9 = \text{"nine"}$; $8 \times 4 = \text{"twenty four"}$). Intrusion errors are important because, as summarized below, they demonstrate notation-dependent interactions of numerical encoding, retrieval, and response mechanisms that cross basic functional boundaries assumed in the modular model. In other words, operand-intrusion errors appear to be products of strikingly nonlocal, interactive processes.

Campbell (1993) demonstrated three prominent characteristics of operand intrusions that support these conclusions. First, intrusions were much more frequent with problems in number-word format (e.g., nine \times six = ?) as opposed to Arabic-digit format ($9 \times 6 = ?$), and the effect of notation on intrusions varied with arithmetic operation (i.e., multiplication or addition; see Campbell, 1993, for a detailed discussion). Notation \times operation interactions demonstrate that some processes of calculation vary with surface form. Second, intrusions frequently preserved the number-word lexical class (i.e., *tens* or *ones* words) corresponding to the input-order or position of the intruding operand (e.g., $6 \times 9 = \text{"sixty three"}$; $9 \times 6 = \text{"thirty six"}$). This characteristic implies that intrusion errors are due to spontaneous number-reading processes that converge with calculation processes at some point in processing. A third feature of intrusions demonstrates that the point of convergence cannot be localized at a "postcalculation" stage: if operand intrusions arose by the priming of postretrieval lexical processes, then intrusions would frequently produce random answers. Instead, intrusion errors usually involved answers that were associatively or semantically related to the problem (e.g., $8 \times 4 = \text{"twenty four"}$). This implies that numeral reading processes penetrate the ongoing arithmetic retrieval process, possibly because both processes compute similar verbal-phonological structures. Taken together, these characteristics of intrusions demonstrate that aspects of number reading and arithmetic are interactive

and determined by surface notation, rather than separate, notation-independent processes.

Interactive functions and selective deficits. The specific features of intrusion errors produced by normal subjects cast doubt on the locality assumption underlying the modular model; consequently, they also cast doubt on the validity of interpreting selective deficits of number processing as direct evidence about the functional architecture. The notation \times operation interactions revealed by intrusion errors violate the locality assumption because such interactions admit the possibility that a disruption of modality-dependent encoding processes could selectively disrupt specific calculation or production processes downstream. Thus, intact performance on encoding "control tasks" does not necessarily eliminate encoding processes as the source of specific calculation or production deficits. Furthermore, the evidence that distinct numerical functions such as number reading and arithmetic-fact retrieval are interpenetrated greatly complicates the functional classification of selective deficits. For example, the evidence that number reading and arithmetic retrieval processes are strongly interactive in normal subjects implies that a reading-related deficit (e.g., a weakened capacity to control or inhibit number-reading processes) could disrupt calculation, while other number-processing skills, including number reading, were relatively spared.

Although it is likely that there are genuinely modular cognitive systems that respect the locality assumption, complex skills such as number reading, comprehension, and calculation probably are not independently localized. Instead, consistent with Farah's conclusions regarding several other cognitive domains, there is evidence in numerical cognition of a substantial degree of modality specificity and functional interactivity that questions the locality assumption.

Discarding locality assumptions: Problems and prospects

Ruth Campbell

Department of Psychology, Goldsmiths College, London SE14 6NW, United Kingdom

Electronic mail: r.campbell@gold.ac.uk

Farah argues that the locality assumption in neuropsychology is not useful for further theoretical understanding of cognitive processes since local damage in complex dynamic systems has effects beyond the site of damage. In some ways this is akin to showing that when a stone is dropped in a pond there are ripples from the epicentre. It is a useful demonstration for those who might deny that complex dynamic systems, like cognitive ones, including brains, have such characteristics. Furthermore, Farah is right to point out that such denial is a common (though not a universal) feature of theorising in cognitive neuropsychology.

However, as the stone analogy might suggest, there are reasons for not abandoning the locality assumption prematurely. Farah suggests that "we . . . lack clear heuristics for selecting models to test." The locality assumption itself, when used properly, can surely offer just such a heuristic. Farah has previously pointed out that one means of delimiting functional subsystems (hence defining the constituent layers of hierarchically arranged systems for simulation) is to investigate patterns of *double dissociation of function in association with patterns of associated disorder*. She has successfully used this tactic to indicate the precise fractures between face processing and reading (Farah 1991). The double dissociation method is strict – for example, the covert-overt processing distinction would not have been a candidate for system-schism since there are no clear cases of absence of covert processing in the presence of overt skill. Furthermore, constellations of associated disorders can give necessary clues for locating the fault-lines between

subsystems. In other words, Farah and her colleagues are already making effective use of locality assumptions, albeit strictly defined locality assumptions, in driving their theoretical simulations. A less radical conclusion of her target article is that some locality assumptions are justified, others are not.

A second reason for not abandoning all locality assumptions is that their utility (or the lack of it) has not yet been fully tested within simulations themselves. A variety of disturbances instantiated in simulations is needed to extend the validity of Farah's claims concerning the nonlocalisation of effects. These include the effects of lesion combinations: when two stones are dropped in the pond the ripples interact in previously unforeseen ways.

Which model is best? Once we accept Farah's general point concerning the importance of simulations, further problems arise, which she recognises. How do we choose between different means of modelling complex dynamic states? For face recognition, Burton et al. (1991) have described a simple interactive activation model which simulates *precisely* the same phenomena as those instantiated in the model described here by Farah et al. (1993). Yet the organising principles of the two models are sufficiently different for it to be clear that they are not isomorphic. The Burton et al. model has no learning algorithm, hidden units, or distributed representations. It does have a further "localisation assumption" in its distinction between "person identity nodes" and other semantic representations.

Since both these models show covert face processing in the absence of overt face knowledge, how are we to know which is better? Perhaps *all* dynamic models that capture probabilistic representations at different levels might work? Even models with organising principles quite different from those involving distributed activation (other combinatorial mathematical models, for instance) might perform the covert processing trick, too. Massaro (1988) has alerted us to the possibility of overpowerful modelling in parallel distributed processing (PDP). It is a warning that should be heeded.

Prospects: Distant effects of a local analysis. One of the most useful functions of Farah's target article is that it allows us to loosen the self-imposed modularity straitjacket. With the old localising assumptions, remediation and rehabilitation for brain damage were distinct and sombre enterprises. A missing or damaged function cannot be replaced, only bypassed. With these assumptions relaxed, more is possible; not least, an explanation of the wide variety of modes of response and recovery to brain damage. And not just brain damage; striking aspects of cognitive variability can be observed in development as well. Greater functional plasticity may be inferred from the assumption of a missing or partial component within a connexionist system than within a more localised one. Congenital sensory loss provides an example. The primary characteristic of people born deaf is that their mastery of cognitive processes that rely on heard language is far more varied than that of hearing people. Deaf people of very similar background, constitution, and intelligence may or may not achieve spoken language, "inner speech," regularity effects in reading and spelling, and so forth (Campbell 1992; Dodd & Murphy 1992). A strictly modular, localising approach has difficulty with this range of achievements, which would not be predicted when the sensory input is too limited to deliver the required distinctions. An interactive systems approach advocates a more relaxed stance, for activity at higher levels would be slowed and possibly skewed by profound lack of sensory discriminanda, but it need not be blocked. Such an approach could be used to find the conditions under which, as in the case of the deaf with good language, "a little can go a long way." The loosened approach advocated by Farah may also help us understand some apparently modular developmental phenomena which are nevertheless not as circumscribed as "tight" modular theory suggests, such as specific language impairment (Bishop 1992) and its apparent converse: language skills in the "savant" with poor intelligence (Smith & Tsimpli 1991). This is a long way from the cognitive neuropsychological focus of Farah's

paper, yet the implications reach out quite directly from it, just as the ripples from the stone cast in the pond.

Casting one's net too widely?

D. P. Carey and A. D. Milner

Psychological Laboratory, University of St. Andrews, St. Andrews
KY16 9JU, Scotland, United Kingdom

Electronic mail: dpc1@st-and.ac.uk; adm@st-and.ac.uk

1. The locality assumption as a starting point. Dissociations are the meat and drink of experimental neuropsychology. But no such observation is strong enough to enable the investigator to make an "inference" from it. To suppose otherwise is to misrepresent most neuropsychological reasoning. In general, the data lead instead to a *hypothesis*, and the first hypothesis to be entertained should be the simplest one that would (a) fit the facts, (b) have some biological plausibility, and (c) be testable. In practice, the initial hypothesis will generally be one implicitly based on at least a weak version of the locality assumption. But a hypothesis is only a beginning, never a conclusion: the next step is to look for independent evidence that might challenge or support the hypothesis. Ideally, one seeks convergent evidence from other forms of enquiry such as neurophysiology or functional anatomy. In many cases in cognitive neuropsychology, however, this may be unobtainable, requiring one to make do with fresh neuropsychological evidence garnered from other patients or from normal subjects.

In different parts of the brain and in different functional systems, the likelihood that some version of the locality assumption might be helpful when framing an initial hypothesis will vary enormously. In the visual system, for example, a great deal of modularity appears to be present (Felleman & Van Essen 1991). Therefore, when a behavioral dissociation is discovered in this domain, hypotheses as to the role of different areas of the cortex do not generally deserve to be greeted with hollow laughter. A recent example is a hypothesis generated to explain a double dissociation between the visual capacities of an agnostic patient (Goodale et al. 1991; Milner et al. 1991) and patients with optic ataxia (Jakobson et al. 1991; Perenin & Vighetto 1988). This hypothesis, that the functional architectures for visual recognition and visuomotor guidance might be largely independent, was strengthened by data from behavioral, neurophysiological, and neuroanatomical findings (Goodale & Milner 1992; Milner & Goodale 1993). In no sense was the proposal *inferred* from the empirical dissociations found in the two types of patient, however: rather, the findings were used along with other available evidence to generate a testable hypothesis that makes explicit predictions about the functional properties of two streams of visual processing in human and nonhuman primates.

That hypothesis (and in fact *any* hypothesis about functional modules in the visual system) has to be constrained by the current evidence for elements of both serial and parallel processing, and for considerable cross-talk between separate cortical areas and the functional modules that may lie within them (Merigan & Maunsell 1993). The modular proposals of Goodale and Milner accordingly have to be tempered with a recognition that the hypothesized functional streams in the cortex would not operate in isolation from one another. In fact, the obvious need for object identity to be able to inform visually guided prehension demands an interaction between the two systems. A patient has recently been described (Sirigu et al., personal communication) who cannot use object knowledge in this way; it may thus be hypothesized that she has suffered a neural disconnection between the two systems.

Although different functional modules will inevitably interact with each other much of the time, this does not mean that in some instances they may not behave rather autonomously.

Those instances may provide the initial evidence for modularity; to ignore such evidence because it appears "naive" to accept the locality assumption would be a dereliction of scientific duty. Of course, the conditions necessary for the occurrence or nonoccurrence of such instances would be incorporated into the functional hypothesis.

2. Nets: What do they catch? The explosion of neural net models in the recent neuroscience and cognition literature reflects the immense fascination these models have for many researchers (e.g., Hinton 1992), but although neural nets often have fascinating properties, in practice many of the proposals for direct analogies to brain/cognitive function can be highly problematic (Crick 1989). In particular, any neural net which produces a desired output from a specified input is hugely underconstrained; an infinitely large number of solutions can be found for each problem addressed (Fodor & Pylyshyn 1988; Reeke & Sporns 1993). That is, solving an input-output problem which has several computable solutions means little more than that the problem is solvable; for such nets to model brain function they have to do more. The explanatory utility of a given net is rather limited unless it has at least two properties. First, it should be biologically plausible; second, it should lead to testable predictions in normal subjects and patients, predictions not specified by the input/output characteristics of the system it purports to model (see Reeke & Sporns 1993).

As a class of models, neural nets undoubtedly provide a step in the right direction insofar as they emphasize plasticity, interconnectedness, and parallelism. The evidence for structures in the real CNS that look like the postulated nets is still relatively scant, however (Eagleson & Carey 1992). All three networks endorsed in the target article (like many others, e.g., Kettner et al. 1993; Plaut & Shallice 1993) utilize the back-propagation algorithm, which has been repeatedly criticized for its lack of biological feasibility (Crick 1989; Eagleson & Carey 1992). Others have made attempts to build more "biologically plausible nets" (e.g., Mazzoni et al. 1991a, 1991b), but these contain similarly questionable assumptions about brain function. For example, the learning rule now advocated by Andersen and his colleagues (Mazzoni et al. 1991a, 1991b) does not bypass the "spatial crosstalk" problem (conflicting error messages to the same hidden unit), which is a difficulty for it and for many other nets (Jacobs & Jordan 1992).

3. Disengagement of visual attention. Last, we wish to question whether the second of the author's three examples can be correctly characterized as an instance of the locality assumption. Impaired shifting of visual attention was experimentally documented in patients exhibiting clinical "extinction" following unilateral damage to the parietal lobe (Posner et al. 1984). The patients had a particular problem in detecting visual signals in the contralesional field following an invalid warning cue located in the ipsilesional field. The authors hypothesized that the deficit was one of disengaging attention from the cue, but the patients were also impaired even when the warning cue was placed centrally (whether it was symbolic or neutral), provided that the target stimulus was contralesional. The only kind of "disengage" deficit that could have explained the impairment accordingly had to be one of disengaging-attention-in-a-contralesional-direction. And indeed more recent evidence directly supports such a directional interpretation (e.g., Posner et al. 1987). Thus, the data were never explicable in terms of a "disengage" operation independent of later components in the attention-shifting process.

If a pure "disengage" operation would not figure in any plausible hypothesis to account for the neuropsychological data, however, how does the particular model proposed by Cohen et al. (in press) help our understanding of this disorder of shifting attention? It certainly does not explain the patients' difficulty in shifting attention from a central site in the contralesional direction. No doubt it could be changed in an *ad hoc* way so that it did, but how does one then choose among the many possible

different neural net models that could be devised? We remain uneasy about the heuristic and explanatory value of a class of theories against which no evidence can ever count decisively.

Modularity, interaction and connectionist neuropsychology

Nick Chater

Neural Networks Research Group, Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom

Electronic mail: nicholas@cogsci.ed.ac.uk

Farah argues that cognitive neuropsychology assumes a modular cognitive architecture, in Fodor's (1983) sense, and that this leads naturally to the "locality assumption." She recommends an alternative class of computational models, interactive connectionist networks, which violate locality. Although the specific interactive connectionist models she discusses are interesting alternatives to existing box-and-arrow accounts in their respective domains, the general arguments they are intended to illustrate are less compelling.

First, violations of locality are common in modular as well as interactive systems. Consider the muscular system, which has a clearly defined modular structure. Damage to one component (for example, straining a particular leg muscle) may cause significant compensatory changes in the behaviour of others (causing a completely different gait, or even a different method of locomotion – e.g., hopping rather than walking). Thus, the behaviour of a component, even in a modular system, may very well change immediately if another component of that system is damaged. In psychological terms, one would say that damage may cause patients to change their strategy for carrying out a particular task. For example, a subject who has lost the putative lexical reading route might start to rely on phonological or semantic routes which were not involved in premorbid reading. Nonetheless, whereas what we might term "behavioural locality" may be violated in such situations, locality of function need not be. The functional capabilities of the individual muscles (i.e., the forces they can generate) will presumably be unchanged immediately after damage elsewhere in the muscular system. However, these functional capabilities will themselves rapidly alter as the system becomes adapted to the new mode of function. Just as muscles adjust rapidly to their new role, so components of a modular cognitive system may rapidly learn to adapt to their new cognitive function. Violations of locality, either behavioural or functional, will make it very complex to draw inferences about normal function from impaired performance.

Second, the modularity thesis (Fodor 1983) is not addressed by Farah's models, despite being the subject of the introductory discussion. Fodor's contention, which Farah opposes, is that the cognitive processes involved in perceptual analysis, motor control, and language processing are organized into modules which are informationally isolated from one another and from the unencapsulated central processes which mediate common sense thought. The precise grain of such modules is not specified, but Fodor's principal concern is to defend the view that large cognitive domains (e.g., language processing, visual analysis, etc.) are subserved by separate modules. This position is entirely consistent with the models that Farah presents: one model concerns memory, which is generally not thought to be informationally encapsulated, and the others can reasonably be interpreted as partial specifications of modules for attention and face recognition. Furthermore, the assumption of some kind of global modularity seems to be a presupposition of the very attempt to model a specific cognitive function. If the functioning of the face-recognition system, say, is really intimately bound up with the function of many or even most other cognitive pro-

cesses then a free-standing face-recognition model is surely not possible.

Third, the emphasis on the interactive nature of connectionist models is idiosyncratic. Although McClelland (1991) emphasizes interaction in his GRAIN networks, most connectionist models are feedforward networks (or variants) trained by back-propagation. In experimental cognitive psychology many of the same phenomena may be captured by both interactive and feedforward network architectures (e.g., McClelland & Elman 1986; Norris 1990; Shillcock et al. 1992). Furthermore, connectionist neuropsychological models, such as Patterson et al.'s (1989) model of surface dyslexia and Hinton and Shallice's (1991) model of deep dyslexia, derive interesting and detailed predictions using feedforward networks. Since the analysis of the general patterns of breakdown observed in even simple feedforward networks is extremely difficult (Bullinaria & Chater 1993), it is surely much too early to decide between alternative network architectures for neuropsychological modelling.

What is fundamental, and what rightly takes centre stage in Farah's general discussion, is the difference between connectionist neuropsychological models and the traditional box-and-arrow approach. Traditional box-and-arrow models are so underspecified that only very gross patterns of damage largely concerning task dissociations can be predicted. [See Précis of Shallice's *From Neuropsychology to Mental Structure*, *BBS* 14(3) 1991.] By contrast, connectionist models are fully specified mechanisms on which the behavioural effects of all manner of damage can readily be tested, and which, when intact, can be assessed as models of normal performance. This is perhaps the real promise of Farah's work and that of the rest of the growing field of connectionist neuropsychology.

ACKNOWLEDGMENT

This work was supported by grant SPC-9029590 from the Joint Councils Initiative in Cognitive Science/HCI.

Modularity, abstractness and the interactive brain

James M. Clark

Department of Psychology, University of Winnipeg, Winnipeg, Manitoba, Canada R3B 2E9

Electronic mail: clark@uwpg02.uwinnipeg.ca

Farah has contested the assumption that brain functioning is localized or modular and has argued for a highly interactive brain. I cite another example against modularity, describe an added benefit of the competing associative view, and challenge further the received view of brain functioning.

Number processing. The locality assumption rejected by Farah for semantic taxonomies, visual attention, and face recognition is also central to other areas. In number processing, McCloskey and his colleagues (e.g., McCloskey et al. 1986; 1992; Sokol et al. 1989) have proposed a modular view based on distinct comprehension, calculation, and production modules that communicate solely by mediating abstract number codes.

Campbell and Clark (1988; 1992; Clark & Campbell 1991) have presented an alternative, encoding-complex view of number processing in which numbers are represented as concrete codes in diverse formats (e.g., digits, number words, analogue codes). In place of function-specific modules, interactive excitatory and inhibitory associations among specific codes perform number identification, calculation, and production.

The arguments advanced against modular views of number processing have reflected criteria similar to those cited by Farah. In particular, nonlocalized associative theories can accommodate findings thought to support modularity and can explain phenomena that are awkward for modular views. The

abstract number codes that segregate modules are also questionable (see below). Although these claims have been challenged (see papers cited earlier), the example nonetheless demonstrates the generality of the issues and arguments advanced by Farah.

Associative models. Modular views are weakened by demonstrations that nonlocalized associative theories can explain behavior in terms of excitatory and inhibitory connections among mental representations. Associative theories include connectionist models, such as those described by Farah, as well as related approaches that do not assume distributed representations (e.g., Campbell & Oliphant 1992). Farah points out the empirical adequacy and other benefits of such models.

One particular strength of associative models not emphasized by Farah is that they are undeniably mechanistic; that is, they identify physical events (e.g., representations, activation) intervening between inputs to and responses of the cognitive system. This mechanistic quality elevates associative models above psychological theories that interpret behavior by abstract symbolic processes (e.g., "if-then" procedures, retrieval) that all too often say little about concrete, underlying mechanisms. The associative approach compels researchers to deal with the underlying mechanisms, or at least to admit their present ignorance about those mechanisms. In turn, the translation of psychological metaphors into physical mechanisms will perform reveal the associative quality of the underlying causal links and neuronal systems.

Associationism has a controversial history. Associative models have been criticized for being vague and weakly specified, and for lacking formal constraints. Farah correctly noted that connectionist models are not intrinsically more *post hoc* than high-level, symbolic models, and also that empirical constraints should be more important than formal constraints. Undue emphasis on formal properties has contributed to the unwarranted faith in modularity and obstructed the development of mechanistic, associative models. Bever et al. (1968), for example, argued on formal grounds that associative models in principle could not explain many facets of human behavior. Such arguments count for little in the face of successful connectionist and other associative models.

The received view. Farah challenged the tacit and widely held assumption that brain and cognitive processes are localized and modular, but the received view is based on other fundamental premises that are similarly doubtful. In particular, a critical evaluation is needed of the assumption that abstract semantic codes and processes underlie human behavior. The abstract code and locality assumptions tend to cooccur (e.g., abstract codes define the boundaries between McCloskey et al.'s modules).

Despite rejecting modularity, Farah retained abstract semantic codes and, implicitly, the assumption of a distinct semantic module. This is clearest in her models for taxonomic categories and face perception (Figs. 1 and 11). Figure 11, for example, identified special semantic units to identify such features as "actor." This abstract code assumption is unnecessary, inasmuch as the word "actor" and other similarly specific codes can subserve functions attributed to semantic codes and can avoid the artificial distinction between semantic and nonsemantic processing modules (i.e., hidden "locality").

Thus Farah unadvisedly left intact a second central fallacy of much cognitive and brain theorizing, namely, that a semantic system exists distinct from patterns of activation in specific verbal or nonverbal codes. According to strong associative views (e.g., Campbell & Clark 1988; Clark & Campbell 1991; Paivio 1986), meanings and concepts emerge from interactive brain processes involving associations among words, objects, motor images, and other concrete representations. The added assumption of abstract, semantic codes is superfluous.

Conclusions. Farah's challenge to locality is a positive step toward ridding the behavioral and brain sciences of unwar-

ranted, restrictive assumptions about the human brain and related psychological processes. The modularity assumption and correlated claims about abstract codes lack a sound empirical foundation. They only became dominant because of fallible rational arguments and because their proposed view of cognition and the brain was amenable to scientists' thought processes and available computational tools. More generally, Farah has demonstrated how vital it is that every scientific assumption, no matter how rational it seems, be questioned, be put to rigorous empirical test, and be challenged by contrasting theories, such as associative models based on interactive excitatory and inhibitory mechanisms.

ACKNOWLEDGMENT

Preparation of this paper was supported by grant OGP0042736 from the Natural Sciences and Engineering Research Council of Canada.

Further advantages of abandoning the locality assumption in face recognition

Jules Davidoff^a and Bernard Renault^b

^aDepartment of Psychology, University of Essex, Colchester CO4 3SQ, England and ^bUniversity of Paris 6, CNRS URA654-LENA, Hôpital de la Salpêtrière, 75651 Paris Cédex 13, France
Electronic mail: ^ajdavid@sx.ac.uk; ^bmartiner@arthur.citi2.fr

Farah proposes that dissociations reported in the neuropsychological literature are the result of damage to interactive neural networks and not to impairments at "localised" sites. There is, in essence, nothing new about her opposition of global and stage accounts of intellectual functions. The long history of the debate can be gathered by Farah's introductory quotation from Ferrier (1886) and it could have been equally replaced by ones from Charcot (1883) or Freud (1891/1935). However, Farah is explicit in her parsimony. For example, she shows how a neural network for faces can be damaged to effect chance levels for overt recognition but still retain connections that can facilitate learning. A consequence of the model is that there is no need to postulate a magic box within which resides the "conscious awareness system" (de Haan et al. 1992).

Farah sidesteps the impossible task of defining consciousness by showing that the results from overt and covert tasks are interpretable from the same implementation; we recommend that the exercise be extended. Network models would find it convenient and more parsimonious if the implementation could also include differences between the processing of unfamiliar and familiar faces. These are unlike local models of the Bruce and Young (1986) type that use a clearly separate stage for the processing of familiar faces.

The study of prosopagnosia has been taken to validate the locality assumption, but the evidence is somewhat in dispute. Not one of the nine prosopagnosics tested by Schweich and Bruyer (in press) showed normal performance in all tasks with unfamiliar faces; similarly for the four prosopagnosic patients of Davidoff and Landis (1990). The frequent connection between impairments on tasks with familiar and unfamiliar faces indicated to Davidoff and Landis (1990) that common processing was involved. They suggested that the intact status of the neural implementation should be assessed with superiority tasks based on those of Homa et al. (1976). These tasks use unfamiliar faces, and prosopagnosics, unlike other patients with posterior brain damage, do not show object superiority effects. To accommodate the Davidoff and Landis (1990) results, a local processing model would be forced to infer that prosopagnosic deficits were based on failure at the structural encoding stage for faces. Farah's model, however, might be extended to cover impairments to unfamiliar faces from a single lesion of the network.

We accordingly applaud the attempt to produce a network implementation for face recognition but still feel it necessary to

ask whether Farah's model provides a satisfactory account for all covert tasks that have been used with prosopagnosic patients:

First, why does the degraded representation produce a GSR (galvanic skin response) to an unrecognised familiar face (Bauer 1984)? Is it because connections to limbic structures may be activated by only part of the face representation? Compare, for example, the equal GSR response to whole and part stimuli demonstrated for words by Fuhrer and Eriksen (1960) in an earlier debate concerning unconscious processing (see Dixon 1981).

Second, would lesions to Farah's network produce the systematic changes to evoked potentials observed by Renault et al. (1989)? They showed, for faces that were not overtly recognised, that the amplitude of the P300 evoked potential was inversely related to the probability of presentation of the face category. Could Farah's model be sensitive to such short-term manipulations of probability? Furthermore, the latency of the P300 ERP – an index of stimulus evaluation time – increased with the familiarity of the face. The results of Renault et al. (1989) imply that extended coding produces an "attractor" from which it takes a long time to escape. Could this be incorporated into Farah's model?

Neurocomputing and modularity

Joachim Diederich

Neurocomputing Research Centre, School of Computing Science, Faculty of Information Technology, Queensland University of Technology, Brisbane Q 4001, Australia

Electronic mail: joachim@fitmail.fit.qut.edu.au

Farah claims that inferences about the functional cognitive architecture can be made from neuropsychological data with a "nonmodular" set of assumptions, that is, human information processing is graded, distributed, and interactive. To support this claim, Farah presents simulation data from three different interactive systems. The nature of these simulation systems is important, because according to Farah (sect. 3.3.2), the microstructure of a cognitive system is relevant (if the locality assumption is abandoned) because of its implications for the macrostructure, that is, the functional cognitive architecture.

Although Farah stresses the importance of microstructure, all three models fail to include common knowledge about it, for example, connectivity patterns in the brain, the elaborate structure of neurons, and so on. In addition, there are a number of biologically unrealistic assumptions: neurons have both excitatory and inhibitory connections; there is no differentiation of neurons in various morphological types or transmitter types, and neurons communicate by exchanging activation values (i.e., there is no coding of spike frequency or other communication types; e.g., hormones).

Farah is using simple associative networks to show that a strongly interactive system can explain established data without the locality assumption. The simple associative networks are inadequate for neural modeling, however, and violate common knowledge about neural processing. It is therefore an open question what kind of inferences can be made based on the simulation results, in particular, what kind of inferences can be made based on damaging biologically unrealistic networks.

It is widely acknowledged in modeling systems that an input-output function at one level of organization can be simulated by an infinity of models based on subcomponents from lower levels (Shepherd 1990, p. 83). For cognitive and neural modeling, this implies that the subcomponents should not be arbitrary but must represent properties which can be found in real biological systems. Farah fails to address this issue but uses a number of assumptions that are inconsistent with common knowledge about neural information processing.

Furthermore, the effect of damage to the networks can easily

be explained by the network structures. For example, in simulation 1, category-specific impairments (sect. 2.1.3), there are seven times as many visual semantic units for living things as there are functional semantic units. Therefore, because most of the semantic input to the name units of living things is from visual semantics, damage to visual semantics eliminates a greater portion of the excitation needed to activate the names of living things (sect. 2.1.3). Farah claims this result contradicts the locality assumption because "when visual semantics is damaged the remaining parts of the system do not continue to function as before." This is not really surprising, given the way the network is set up.

Farah (following Shallice 1988) points out that modularity compared to interactionism is a matter of degree. A strong point for interactionism can certainly be made, but Farah fails to address a number of issues that are important for a highly distributed, interactive system. For example, if a conceptual representation is spread over some fraction of units, the chances are high that the encoding will overlap and cause cross-talk (Feldman et al. 1988). The biggest problem, however, is the communication between subnetworks: because of the distributed representation, only one "concept" at a time can be transmitted between subsystems if each concept is a pattern on the connections between the subsystems. Such a system is therefore highly sequential because computation is based on passing simple (numeric) messages. This communication problem does not appear in Farah's relatively simple simulations but will be relevant as soon as bigger networks with a large number of subsystems are considered.

Work in computational neuroscience suggests better, alternative ways of modeling: unsupervised or reinforcement learning in sparsely connected networks, compact but still distributed coding and communication via random links between subsystems. Modeling that is constrained by recent results from experimental neuroscience can make a stronger point for interactionism than Farah's simulations.

ACKNOWLEDGMENT

Thanks to Shlomo Geva for his comments on earlier drafts of this commentary.

Clarifying the locality assumption

Clark Glymour

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213

Electronic mail: clark.glymour@andrew.cmu.edu

1. Distinctions. Farah's arguments turn on ambiguities in the meaning of the "locality assumption." It is easy to state the alternatives formally. A component *C* has input channels I_1, I_2, \dots, I_k from other components and possibly input channels I_{k+1}, \dots, I_{k+m} from the environment, where each channel is viewed as a variable that can take on any of some range of values. Component *C* likewise has output channels O_1, \dots, O_j to other components, and perhaps output channels O_{j+1}, \dots, O_r to the environment. In normal functioning, each component thus determines an input/output function: $\mathbf{O} = F_C(I_1, I_2, \dots, I_{k+1}, \dots, I_{k+m})$, where \mathbf{O} is the vector of outputs. Ordinarily, the set of inputs and outputs are assumed to be disjoint, but if F_C is understood to involve a time delay, however short, the sets of input and output channels need not be disjoint. Since the inputs I_1, I_2, \dots, I_k are from other components, whose outputs are functions of still other components and of the environment, and so on, if we hold constant any entirely internal variables, the value of \mathbf{O} is also a function of some larger set of environmental inputs, $\mathbf{O} = H_C(E_1, \dots, E_p)$, where some of the *E* variables are I_{k+1}, \dots, I_{k+m} . When the system is not normal because some component *D* other than *C* is damaged, there are corresponding input/output functions for component *C*, which we can denote

by $F_{CID}(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m})$ and $H_{CID}(E_1, \dots, E_p)$ respectively. Here are several logically distinct locality assumptions:

- (1) For any two distinct components, *C* and *D*,

$$F_C(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m}) = F_{CID}(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m})$$

- (2) For any two distinct components, *C* and *D*,

$$H_C(E_1, \dots, E_p) = H_{CID}(E_1, \dots, E_p)$$

- (3) For two particular distinct components, *C* and *D*,

$$F_C(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m}) = F_{CID}(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m})$$

- (4) For two particular distinct components, *C* and *D*,

$$H_C(E_1, \dots, E_p) = H_{CID}(E_1, \dots, E_p)$$

- (5) For a particular component *C* and for all distinct components *D*,

$$F_C(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m}) = F_{CID}(I_1, I_2, \dots, I_k, I_{k+1}, \dots, I_{k+m})$$

- (6) For a particular component *C* and for all distinct components *D*,

$$H_C(E_1, \dots, E_p) = H_{CID}(E_1, \dots, E_p)$$

I do not know which of these claims psychologists intend; we rarely make ourselves clear. But one would be foolish to endorse (2), which would be equivalent to thinking that even though the battery is dead the car motor will turn the crankshaft which will turn the wheels which will make the car move when you switch on the ignition key. (6) may be a consequence of what is meant by the claim that module *C* is "informationally encapsulated." Claims of the form (3), (4), (5), or (6) might be of scientific interest if particular components could be identified. In the absence of clarity, it would be charitable to suppose that (1) is what is usually meant by the locality assumption in general, and that (3) is what is meant when people speak of locality for particular modules. In any case, since (2) is so obviously and trivially false, it has no scientific interest.

Unfortunately, Farah's examples only argue against (2). The first of her examples gives a connectionist model in which various disjoint collections of nodes are labeled as modular components: visual semantics (*vs*), functional semantics (*fs*), vision (*v*), and names (*n*). In the context of the model the two latter serve as environmental inputs, so that normally,

$$fs = F_{fs}(vs, v, n); vs = F_{vs}(fs, v, n)$$

Farah lesions *vs* and finds that,

Contrary to the locality assumption, when visual semantics is damaged the remaining parts of the system do not continue to function as before. In particular, functional semantics, which is part of the nondamaged residual system, becomes impaired in its ability to achieve the correct patterns of activation when given input from vision or language. This is because of the loss of collateral support from visual semantics.

What she has found is that (4) is not true when $C = fs$ and $D = vs$ and the *E* variables are *v* and *n*. Now of itself this is not of much interest, since no one had contrary views about *these particular* modules, and since the falsity of (4) for this case is obvious from the functional structure. If the case is a counterexample to some more general thesis, that more general thesis can only be (2), since the example has no bearing at all on (1). And because it is trivially false, (2) is of no interest.

The same considerations apply to the second example, which provides a model of visual attention without a "disengage" module. I do not object to the model Farah develops, but its behavior when lesioned is only a counterexample to the absurd localization thesis (2), and not to anything more interesting.

Farah's third example, concerning relearning face recognition after damage, is even more remote from any clear, interesting general thesis about functional localization. It does show an important capacity of neural net models to account for learning phenomena that box-and-arrow diagrams do not address at all, and that fact, not confusions about localization, ought to be the lesson readers take from it.

Farah's models are interesting in themselves, and may even be correct, or at least more correct than the alternatives she discusses. I think she is also entirely right that connectionist architectures form an enlightening set of models to account for the data of cognitive neuropsychology. But the central line of argument in the target article is very unfortunate.

2. What are functional modules? Farah's paper prompts a question cognitive psychologists ought to try to be clear about: *In a connectionist architecture, what is a functional module or component?* A principal aim of the contemporary revival of neuropsychology has been to get information about what modules there are, what they do, and how they are or are not connected. But if a connectionist architecture is correct, what is a functional module?

One might say: a functional module is any collection of nodes with an input set that is not the set of all other nodes and an output set that is not the set of all other nodes. That proposal allows distinct modules to share a subset of their nodes, which seems reasonable, and keeps the function in functional. But then what are the functional modules in a network in which every node is adjacent to every other node (save possibly that environmental nodes may not be connected directly to one another)? For any set of internal nodes will receive input from every other node and give output to every other node. Under this definition, such a system has a single functional module.

Perhaps one should say instead that a functional module is any set of nodes that form a clique – a set in which every node is adjacent to every other node in that same set. That definition also permits distinct functional modules to have a common subset of nodes and preserves the formal content of "functional." Then in a completely connected network we get a conclusion opposite to the one in the previous paragraph: in a completely connected network, every subset of nodes forms a functional module.

Perhaps we should say that a functional module isn't specified by input and output connections but by some feature of the aggregate probabilities or strengths of influence a set of nodes establishes between two other sets of nodes, an aggregate which will be determined as much by the weights attached to links as by the network topology. But then the notion of a functional module seems to become vague, for strengths of influence and probabilities are matters of degree.

What exactly can be determined about functional modules in connectionist networks from the environmental input/output behavior of normal and lesioned systems will depend on which of these – or other – senses of "functional module" are adopted. Until that is made clear, the combination of cognitive neuropsychological data and terminology with connectionist computational models offers endless possibilities for confusion.

And, finally, perhaps if the true architecture of cognition is connectionist, we should stop talking about functional modules altogether and just talk about networks and their topologies and weights.

No threat to modularity

Yosef Grodzinsky and Uri Hadar

Department of Psychology, Tel Aviv University, Tel Aviv 69978, Israel
Electronic mail: yosef1@ccsg.tau.ac.il; uri-h@ccsg.tau.ac.il

In what aims to be a neuropsychological salute to connectionism, Farah presents an argument with the following structure:

1. There is a ubiquitous assumption in neuropsychology regarding the "locality" of functional lesions.

2. There is a family of models – of the connectionist variety – that violate this assumption.

3. Neuropsychological data are compatible with the latter models.

4. Hence, the locality assumption is "probably not correct," and since it is the most basic tenet of the class of models standardly assumed, this class of models is "dubious."

This argument contains three premises and a conclusion. Of the premises one is probably correct (2), two are incorrect (1 and 3), and the conclusion is a non sequitur (4).

1. Consider, first, the locality assumption as Farah sees it. In a modular system, she contends, "each component minds its own business and knows nothing about most of the other components. What follows for a damaged system is that most of the components will be oblivious to the loss of any one, carrying on precisely as before." Hence, "selective deficit in ability A implies a component of the functional architecture [of cognition] dedicated to A." This is a view of locality all right, but we do not believe there is even one "modularist" who would subscribe to it. Here is why.

Modules are connected. You can, after all, talk about what you see or hear, and make the connection between the smell of things and the way they look. Modularity, in its strictest sense, requires that a module be unable to interfere with the action of another, or gain access to its knowledge base. Yet the very essence of the thesis is that while cross-modular exchange of *instructions* is illicit, exchange of *information* (in one direction at least) is a must. After all, what the system does is manipulate symbols: accept representations as input, transform them through a set of formal operations, and transmit them onward. There may be many modules, yet they live in just one head. Thus, the output of one module is input to the next one down the line.

Returning now to neuropsychology, if one component is damaged, its output (or lack thereof) will reflect the impairment. Subsequent, unaffected modules will be fed abnormal input and will, as a consequence, produce deficient outputs resulting, potentially, in aberrant behaviors that may not be immediately traceable to the damaged part. This is what makes work in neuropsychology interesting and, alas, quite difficult. Modules might thus be "oblivious" to the disruption of others, attending to their business as usual; but clearly, if their input is now irregular, so will be, most likely, the representations they produce.

Farah makes no distinction between the action of a module and the kind and quality of representations it spews out; and in neglecting to do so, she misses the whole point about modular systems. Yet the view expressed here is, so far as we know, the standard view of modularity: a tiny disruption to one module can, in principle, result in a huge impairment to the system as a whole, precisely because each module works on its own; and while its operation is insensitive to that of others, its output is extremely sensitive to the results of the action of others, sensitive, in other words, to input representations. This of course runs contrary to Farah's claim, which confuses the separation of actions – "informational encapsulation" – with links established through information flow. It is for this reason that she interprets data from dyslexia and aphasia in such an odd way. Phonological dyslexia, for example, features a selective deficit in reading nonwords. Under Farah's locality assumption, this would mean there is a special processing component in every head, dedicated to reading such sequences – hardly a plausible claim. Yet the standard view is, of course, quite different: the grapheme-phoneme route serves to explain both phonological and surface dyslexia and is not custom-built for reading nonwords.

Another good example of a modular approach to dyslexia is that of Shankweiler and Crain (1986), who focus on developmental reading disorders. Seeking to steer clear from *ad hoc* claims,

they show how a large body of data from dyslexia – phonological, syntactic, and other deficits observed in developmental dyslexia – is accounted for by postulating a small bottleneck in early phonological processing. Because this deficit occurs so early, it projects all over, with detrimental effects on virtually every part of the language-processing system that happens to be located later. This model is explicitly modular, yet its strength and generality derive from the fact that it does not assume locality the way Farah does.

Similarly, one can demonstrate how the postulation of a rather limited deficit in one type of syntactic representation in agrammatic aphasic patients can account for a wide variety of well-documented comprehension deficits in patients suffering from this condition (see Grodzinsky 1990). Here Farah-style locality shows up in its full weakness. She identifies a processing component with each “ability,” yet she does not specify what an ability is. In the context of language comprehension, her locality assumption would mean either that there must be a dedicated processing component for the comprehension of sentences of every structural type – a claim that is highly implausible for anyone familiar with linguistic considerations – or that the data from agrammatism (patterns of impairment and sparing that are best described in syntactic terms) are unpredictable. Thus, the locality assumption Farah argues against is neither held by anyone, nor is it tenable when empirical considerations are taken into account.

2. With Farah’s second premise – that connectionist models violate the locality assumption – we have no quarrel.

3. Farah’s third claim, regarding compatibility of “neuropsychological data” with connectionist theory, is made as if all cognitive domains were alike (at least with respect to modularity): Thus, demonstrations such as Farah’s (that data from a couple of domains are compatible with them) can “provide existence proofs that principled inferences can be made in cognitive neuropsychology without the locality assumption.” This is far from being the case, however, and claims regarding certain domains cannot be carried over to all others (or even to some central ones). The particular examples presented by Farah come, ironically, from domains which on most accounts are not encapsulated, hence nonmodular, since they appeal to knowledge of the world, attention, and awareness. Yet for other domains such as language (a rather central feature of the human cognitive system), no such compatibility has ever been shown. This has been discussed almost ad nauseam in the literature and thus the arguments need not be reiterated here. Yet it is in such domains that one should seek a demonstration of the type Farah would like to argue for. We have yet to see how such models handle selective language deficits such as agrammatism, in which, as we said before, particular aspects of the syntactic analysis of a sentence are disrupted, resulting in abnormal comprehension patterns. We are unaware of serious proposals of this type.

4. Finally, we get to the non sequitur. It is a point of logic that even if neuropsychological data are expressible in connectionist models, this means little in itself. Theories, as everyone knows, are vastly underdetermined by data, and the compatibility of the available data with one of them does not, unfortunately, deny its potential fit with another. Farah’s attempt to turn a seemingly positive point (the fit between connectionist theories and her data) into a negative one (lack of fit with modular theories) thus fails.

So, modularists need not worry: an attack based on a questionable argument and doubtful premises does not put anything of consequence at risk.

ACKNOWLEDGMENT

The preparation of this manuscript was supported by NIDCD grant DC-00081 to the Aphasia Research Center, Boston University School of Medicine.

Go with the flow but mind the details

Glyn W. Humphreys and M. Jane Riddoch

Cognitive Science Research Centre, School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

The idea that neuropsychological disturbances of cognition can be interpreted in terms of models of normal cognition has a long history and has been one of the driving forces behind the success of cognitive neuropsychology for the past 20 years or so. Farah takes issue with one assumption common to the application of this idea, namely, that neuropsychological disturbances reflect disruption to a single component of the normal cognitive system, allowing other parts of the system still to function intact. Following this “locality” assumption, a patient’s deficit may be used to define the nature of the (now impaired) functional component in the normal processing system (see, e.g., Coltheart 1984 for an example of this argument).

Farah raises two questions: (1) Is the locality assumption wrong, perhaps even positively misleading? (2) Can the discipline of cognitive neuropsychology still make progress if the assumption is abandoned? These questions are answered in the affirmative, and we agree. However, we also wish to add cautionary notes concerning the details of the arguments made in the target article. To “go with the flow,” ignoring cautionary details, can threaten the baby when the bathwater is dispensed.

As evidence against the locality assumption, three pieces of research are cited, all relying on simulation of a neuropsychological disorder in a connectionist model. The deficits concern: apparent category-specific disorders of semantic knowledge, problems in attentional disengagement, and covert knowledge in prosopagnosia. In the first example, Farah argues that modality-specific disorders of semantic knowledge can emerge from a distributed memory system in which different forms of semantic knowledge (visual/perceptual and functional knowledge about objects) are highly interconnected and hence not functionally independent. The locality assumption is wrong because it leads to the inference that semantic knowledge is separated into functionally separable processing systems (e.g., one concerned with visual/perceptual knowledge, one with functional knowledge). In the second example, a simple connectionist model with competitive attentional feedback systems in each “hemisphere” is shown to mimic attention “disengagement” problems when lesioned. Here the locality assumption can lead to the inference that there is a disengagement module in the attention system; the simulation shows that such a module is not necessary. The third example shows that covert recognition can emerge after partial damage to a recognition system (see Shallice & Shaffran 1986 for an earlier version of this account); covert recognition does not necessarily result from the disconnection of intact recognition from a system involved with conscious awareness of recognition. The locality assumption here might lead to the inference of such a disconnection, and hence (unnecessarily) to the inference that a specific system for conscious awareness exists.

The difficulties for the locality assumption demonstrated by these examples emerge from inferences about normal processing based on functional deficits in neuropsychological patients. The dangers in this have been stated before (see Seidenberg 1988 for one example), and the simulations reported by Farah give the critical argument concrete force. Although, as noted by Farah, many cognitive neuropsychologists have been circumspect in their inferences concerning normal performance, and although many have noted that patients need not have deficits functionally localised to a single processing mechanism (e.g., Sergent 1987; Shallice 1988), such qualifying details have too often been ignored.

Concerning Farah’s specific examples, however, several cautionary points can be raised in turn. First, consider the simulation of category-specific deficits. To produce such deficits, Farah

and McClelland (1991) selectively lesioned either the visual/perceptual or the functional-knowledge units in their model. Now, for this to happen in a real brain, these units would need to be anatomically modular even if they are functionally nonindependent because of cross connections. All is well providing the anatomical and functional accounts are independent, but this is likely not to be so. In real neural systems there is a preference for short anatomical connections (e.g., Cowey 1985), and this may in turn produce functional consequences: only systems that are anatomically proximal may develop functional interconnections (see Jacobs & Jordan 1992). Case studies also show that patients with category-specific recognition deficits can have intact visual/perceptual or functional knowledge (e.g., Riddoch & Humphreys 1987; 1992; Sheridan & Humphreys 1993). Such dissociations are not comfortable for a fully interconnected distributed processing system. The distinction between anatomical modularity and functional nonmodularity can be questioned.

Second, consider the case of attentional disengagement. Like Farah, we strongly suspect that disengagement describes a functional operation rather than a distinct processing mechanism in the brain (Humphreys & Riddoch 1993). It may nevertheless be wrong to conclude that only one mechanism is involved when people engage and disengage attention on objects. A simple single-mechanism model such as the one implemented by Cohen et al. (in press) cannot easily explain patterns of dissociation in which one patient shows good orienting of attention to signals (providing that attention is not previously engaged on an object) but gross impairments in shifting attention once engaged, whereas another patient shows the opposite pattern of severity (see Humphreys & Riddoch 1993 for preliminary data). We suggest that multiple mechanisms will ultimately be needed.

Third, consider covert recognition in prosopagnosia. Recent research suggests that there is no simple correspondence between the severity of a recognition deficit and whether a patient manifests covert recognition (McNeil & Warrington 1991), yet this correspondence might be supposed if there were a direct relation between partial damage and covert recognition, as in the Farah et al. (1993) simulation. Again, inferences concerning the validity of alternative accounts should be cautious until all aspects of the relevant data set can be accommodated.

Our three cautionary remarks indicate only that moves towards a complete understanding of both normal and disordered cognition are fraught with difficulties for both interactive and modular accounts. As applied to some disorders the locality assumption has clearly been an oversimplification and may cloud understanding of the underlying mechanisms, but applied to others it may remain valid. What is surely right is that we should look to convergence between approaches to provide fuller understanding and that this understanding should include both computational modelling and due attention to anatomical constraints.

Do neuropsychologists think in terms of interactive models?

Marcel Kinsbourne

Center for Cognitive Studies, Tufts University, Medford, MA 02155

Farah cites three of her published explanations of neuropsychological dissociations to illustrate how productive it is to violate what she calls the "locality assumption" and to promote parallel distributed processing (PDP) simulations in clarifying the mechanisms of such effects. All three are excellent ideas, but they do not exemplify a new approach, and the locality assumption is not a principled guideline for neuropsychology but more of a label

for unimaginative interpretation. I will discuss these points in order.

1. Category specific semantic memory. Warrington suggested that selective difficulties in naming animals or objects are surface manifestations of differential damage to underlying mechanisms that represent knowledge in sensory and functional terms, respectively. Unfortunately, the patients who failed to name animals could not do so through the functional route as well as the visual, and this appeared to disconfirm the hypothesis. Farah accomplishes an ingenious rescue by adding a postulate: if, in a network, a representation is mostly activated through one route and only slightly through another, the latter would not in itself suffice to activate it. Animals are mostly characterized visually, only somewhat functionally; so the naming failure by both routes is explained.

This idea assumes that the lesion is confined to one of the two relevant territories in such a patient. But brain damage does not respect functional boundaries, and from the fact that category-specific anomias are rare we infer that the hypothesized access routes to naming must be close together. It is therefore more than likely than an extensive lesion of one territory to some extent overlaps the other. Assume that more functioning territory is needed to activate a name representation through an unusual than a usual route. This yields the situation in which the subject cannot name animals through the functional route but continues to be able to name objects through that route. No doubt this outcome could easily be simulated.

2. Disengage deficit. Farah models Posner et al.'s (1984) disengage deficit through mutually inhibitory interaction between two sets of attentional units relative to the two target locations in Posner's paradigm and derives a difficulty in disengagement without needing to postulate a separate disengage operation. Her model is an incomplete rendering of the well-known orientational bias model (Kinsbourne 1970; 1987), which explains neglect symptomatology of the type studied by Posner in terms of an imbalance between opponent processors subserving lateral orientation in opposite directions. In a recent review (Kinsbourne, in press), I listed many laboratories that have tested this model in various ways over the years. None of us neuropsychologists knew that we were infringing our alleged locality assumption.

3. Unconscious priming of face identity in prosopagnosia. Farah presents a face-recognition network which, when injured, yields various priming effects for known faces, but not their names, simulating certain prosopagnosics. Assuming that naming is a conscious act, this is an excellent application of the general principle that for the contents of a representation to enter consciousness, it need not be moved to some privileged place (Kinsbourne 1988). The idea is that the depleted but pretrained network retains biases in favor of previously experienced patterns. This suggests the following prediction: there will be no such dissociation between priming and naming for face-name pairs repeatedly presented *after* the injury.

4. Modularity. Fodor (a philosopher, not a neuropsychologist) only hypothesized impenetrable modules for a small minority of the cognitive work of the brain. No one thinks that all specialized areas of cortex are "impenetrable." Everyone admits that the "modules" interact. The trick is not to overcome some philosophic bias, but to determine how modules do it.

5. Locality assumption. Taking a surface dissociation and pinning it on the brain is something I would call "neurologizing." It is not a principled way of dealing with neuropsychological data but a bad habit. At least since Hughlings Jackson in the late nineteenth century, neuropsychologists have included interactive mechanisms in their theorizing.

6. Is PDP modeling being oversold? Neuropsychologists derive their good ideas from many sources and we learn that Farah recently got some good ideas from principles of PDP modeling. But in each of the cases cited, the idea does all the work. The

hypothesized mechanisms are so underconstrained that a simulation could hardly fail.

Network theories of brain functioning have been around a long time (e.g., Pribram 1971). The PDP method does promise to be a nice way of demonstrating their usefulness, particularly if the simulation predicts an effect quantitatively as well as qualitatively. For that to be determined we need better constrained theories than we have at this time.

Neuropsychology: Going loco?

Rosaleen A. McCarthy

Cognitive Neuropsychology Unit, Department of Experimental Psychology, Cambridge University, Cambridge CB2 3EB, United Kingdom
 Electronic mail: mm107@phx.cam.ac.uk

The locality assumption (LA), as presented in Farah's challenging paper, consists of a cluster of thematically related ideas. As articulated, the total set of LA postulates provides rather insecure grounds for drawing neuroscientific inferences. However, it is unlikely that any neuroscientist since the early phrenologists can be blamed for subscribing to the entire LA set (or at least at any one time). Certain aspects of the LA are, as the author states, highly problematic; certain aspects are also ubiquitous in cognitive neuropsychology, but the problematic and the ubiquitous are not all of a piece. I wish to suggest that LA can be unpacked into a number of dissociable and distinct ideas: some of these are clearly more tenable and plausible than others.

A very basic notion of locality underpins the analysis of cognitive function in separate domains such as spoken word comprehension and face recognition. Independent though genetically related models are widely assumed to be appropriate in these cases (e.g., Bruce & Young 1986; Figs. 2 and 11 of the target article). It is agreed that partitioning the problem space of cognition into local sectors – whether on the basis of a priori cognitive theory or even folk-psychological prejudice – is a useful way of dividing labour. Nevertheless, independent models do not necessarily imply independent signs and symptoms in patients with neuropsychological deficit.

To take but one example, problems with face identification are reasonably common in patients whose semantic knowledge of animals is impaired. Of course, it would be possible to account for both deficits and even their occasional dissociation within a single network. (One might even venture to do so by having more "visual" units associated with faces than animals in a model such as Farah and McClelland's [1991]). However, cooccurrent symptoms, whether for faces-plus-animals, or visual-plus-functional attributes of animals, are as likely to be based on anatomical proximity of neural substrata as on cognitive relatedness. Partitioning the problem-space into domains can direct neurocognitive theorising in a way that avoids the traps of argument from association (see e.g., Shallice 1988 for discussion). This involves a degree of LA – but arguably to good effect.

At a finer and even less contentious level, LA represents a variant of the cognitive information-processing tenet that complex skills are made up of more specific elementary processing components. The common ground between this level of the LA framework and the position expressed in the target article is illustrated by Farah's adoption of box-and-arrow conventions (e.g., a separation is made between the domains of "vision" and "semantics" in Fig. 2 and between "face input" and "semantics" in Fig. 11). These models imply that word-picture matching and face recognition are not performed by way of some undifferentiated distributed net, but rather through the interplay of distinct processing subdomains. If such componential information-processing frameworks have any empirical psychological valid-

ity then some set of independent variables must selectively compromise their subcomponents. If these independent variables happen to be based on pathology (rather than on cognitive variables such as word frequency, semantic relatedness, or visual degradation) then we are working with the weak form of LA discussed by Shallice (1988) and McCarthy and Warrington (1990).

In section 1.3, Farah points out that interactions between neural subsystems may lead to higher-order effects and so constrain inference about the organisation of cognitive function. This concern, which is closely related to the issue of subtraction (i.e., that the damaged system may be viewed as the normal system, functioning normally, minus one or more components) is critically dependent on theoretically informed empirical research. The validity of the methodological paradigm of subtractivity is an empirical problem for an experimental science. It seems no more soluble by referring to parallel distributed processing (PDP) models in the 1990s than it was by referring to valve radios in the 1960s (Gregory 1961; Weiskrantz 1968). Subtraction and its corollary, statistical independence, are not assumptions: they are empirical hypotheses. There is ample converging empirical evidence (behavioural, radiological, and neuropsychological) that attests to the utility of the subtractive hypothesis when applied to areas as diverse as object recognition, face recognition, language, reading, writing, arithmetic, and memory (e.g., Ellis & Young 1988; McCarthy & Warrington 1990). Of course, it may be inadequate for other domains, but that is an empirical issue.

What can be problematic is any assumption that brain injury invariably causes an uncomplicated (lower order) correspondence between neurological and cognitive levels of organisation (see sect. 1.2). This is the problem of inferring function from deficit: this trap can be seductive but is conceptually quite distinct from using subtraction as a methodological paradigm or empirical framework. The subtractive methodology is related to the use of factorial experimental designs and information-processing models. Inferences from deficit are a set of more-or-less naive conclusions based on data.

LA is accused of holding that there are direct transparent relationships between neural systems, neurological symptoms, and functional architecture (sects. 1.1 and 1.2). It is this aspect of the LA hypothesis that is the least defensible. However, this aspect appears to be one that the target article (albeit in its more optimistic moments) finds reasonably congenial. The sensory-functional hypothesis of semantic deficits is "more in keeping with what we already know about brain organization" than is a theory based on semantic or taxonomic variables (e.g., Caramazza et al. 1990). This may be the case, but it may also be irrelevant: the theories are cast at different levels and one may be explicable in terms of the other. Similarly, the fact that the brain is a complex biological system that shows nonlocal effects when injured (sect. 3.3.3) does not necessarily entail that cognitive theories must mirror this interactivity. (For example, in the case of blood flow, hyperperfusion in the region of an infarct does not mean that the damaged areas of brain are contributing more to the cognitive performance of the system.) Locality at the cognitive level need entail organology at the neural level (even with encapsulated input modules; Fodor 1983).

What about the target article's case for nonlocal cognitive effects arising from local damage? The inference that a particular conjunction of deficits is an inevitable consequence of the interaction between processing components is open to direct empirical evaluation. At this "micro" level there are already a few challenges to the interactive position. For example, Hart and Gordon's (1992) documentation of the selective loss of knowledge of the visual attributes of animals appears to challenge the Farah and McClelland (1992) model of semantic memory; word-finding difficulties may be unaffected by visual

complexity (e.g., Hatfield et al. 1977), contrary to Tippet and Farah (1992); and Patterson et al.'s (1989) model of phonological reading fails to account for the preservation of nonword reading in surface dyslexia (McCarthy & Warrington 1986). These counterexamples are far from being exhaustive. However, even a comprehensive list would not undermine the enterprise of computational or PDP modeling – any more than a failure to demonstrate local effects would invalidate the use of labeled graphs (a.k.a. boxes and arrows). Models can be wrong and that is why we use them. Their power lies in their empirical vulnerability and the extent to which hypotheses, such as those aired in the target article, can be disengaged from speculative or metaphysical assumptions.

If the target article were correct (in its more pessimistic moments), the task of bootstrapping cognitive theory from neuropsychology would be a pretty tall order. Everything interacts, so the chances of finding any systematic patterns would be limited. What is remarkable, therefore, is the degree of consistency and coherence of the overall enterprise of cognitive neuropsychology/neuroscience. Very substantial achievements have been made over a short time. We have a theoretical understanding of the profiles of breakdown in cerebral injury that correspond to analyses of normal cognitive performance. Furthermore, approximately similar deficits are associated with damage to the same regions of the brain and arise in populations with differing languages, cultures, and learning histories. More recently, neural structure and cognitive function in the intact brain have been brought into register through PET scanning techniques – and the findings seem to correspond with what we have learned from studies of brain-injured subjects. Finally, in some cases remediation can even be guided by the frameworks of cognitive theory. This brief summary of achievements does not look like the curriculum vitae of a doomed or fundamentally misconceived enterprise; fortunately, it seems more like the prospectus for an exciting and energetic science.

Distributed locality and large-scale neurocognitive networks

M.-Marsel Mesulam

Department of Neurology, Beth Israel Hospital, Harvard Medical School, Boston, MA 02215

The field of cognitive neuroscience is supposed to be engaged in an epic struggle between the forces of localizationism (or centrism) and those of equipotentiality (or holism). Farah's target article addresses this struggle. According to Farah, there is a "locality assumption" in cognitive neuropsychology which implies that damage to one component of a functional architecture leaves the other components intact. "In such an architecture, each component minds its own business and knows nothing about most of the other components. What follows for a damaged system is that most of the components will be oblivious to the loss of any one, carrying on precisely as before." Farah argues that such a locality assumption is untenable. I agree. In this commentary, I summarize some observations related to the neurology of hemispatial neglect in order to outline an alternative view of functional localization.

Centrist views of functional localization postulate the existence of a nearly one-to-one relationship between lesion site and behavioral deficit. Clinical observations, however, show that a specific deficit such as hemispatial inattention (neglect) can arise following damage to a number of different cortical and subcortical regions of the brain. Cortical lesions that consistently yield hemispatial neglect, for example, have been encountered not only in the posterior parietal cortex, but also in premotor-prefrontal cortex and the cingulate gyrus. Neuroanatomical experiments show that the core cytoarchitectonic entities of

these three regions (area PG, frontal eye fields [FEF], and areas 23–24 of the cingulate gyrus) are linked to each other by extensive and reciprocal monosynaptic connections. The additional subcortical areas where lesions are known to cause neglect (the superior colliculus, striatum, and the thalamic pulvinar nucleus) are connected to at least two of these three cortical foci. These considerations have led to the suggestion that spatial attention is subserved by a distributed large-scale network with three cortical components (or local networks), each providing a slightly different coordinate system for mapping the environment (Mesulam 1981).

Experimental observations suggest that the FEF and area PG may have a collective mechanism for specifying whether a location in space (and events within it) will become the target of enhanced neuronal effects, visual grasp, manipulation, or exploration. In the most figurative and anthropomorphic sense, it could be said that area PG sculpts the subjective attentional landscape, while the FEF and surrounding areas plan the strategy for navigating it. The role of the cingulate gyrus is the least understood aspect of this network. The pattern of neural connections suggests that PG and FEF receive information about the motivational (or limbic) relevance of sensory stimuli mostly through inputs from the cingulate region. The cingulate component could thus introduce a value system into the perceptuo-motor mapping of the extrapersonal space.

Implicit in the preceding account is a dichotomy between the sensory and motor components of directed attention, coordinated respectively by PG and FEF. These two areas are so tightly interconnected, however, that such dichotomies between action and perception become blurred. Although the affiliations of area PG are mostly sensory and those of FEF mostly motor, area PG also contains neurons that fire in association with saccades and reaching movements, and the FEF region contains neurons with well-defined receptive fields (Goldberg & Segraves 1987; Lynch et al. 1977). From a behavioral point of view, a sensory representation is necessary for the accurate guidance of exploratory movements just as exploratory movements are necessary for realigning sensory receptors and updating perceptual representations.

Although clinical observations show that motor and sensory components of unilateral neglect are differentially affected after frontal or parietal lesions, respectively, the dichotomy is not absolute (Daffner et al. 1990). What we see is not the isolated disruption of one or the other behavioral component but the relative salience of one over the other depending on the site of the lesion. The presence of tight interconnectivity between FEF and PG also leads to the expectation that damage to one of these two sites should induce physiological dysfunction in the other, an expectation that has recently been confirmed (Fiorelli et al. 1991). This phenomenon of "diaschisis," known to neurology for more than 50 years, provides one of the many good reasons for dismissing the locality assumption.

Each of the three cortical components in the attentional network serves a dual purpose; that is, it provides a local network for regional neural computations and a nodal point for the convergence and reentrant accessing of distributed information. All three core components are probably engaged simultaneously and interactively by attentional tasks and it is unlikely that there is a temporal or processing-level hierarchy among them. The resultant phenomenon of directed attention is not the sequential sum of perception plus motivation plus exploration but an emergent (i.e., relational) quality of the network as a whole.

A central feature of networks is the absence of a one-to-one correspondence between anatomical site, neural computation, and complex behavior. This is shown in Figure 1. Let behavior alpha correspond to directed spatial attention. Its three major neural computations, A1, A2, and A3, are distributed in sites I, II, and III, which correspond to the FEF, area PG, and the cingulate gyrus. Most but not all of computation A1 is performed

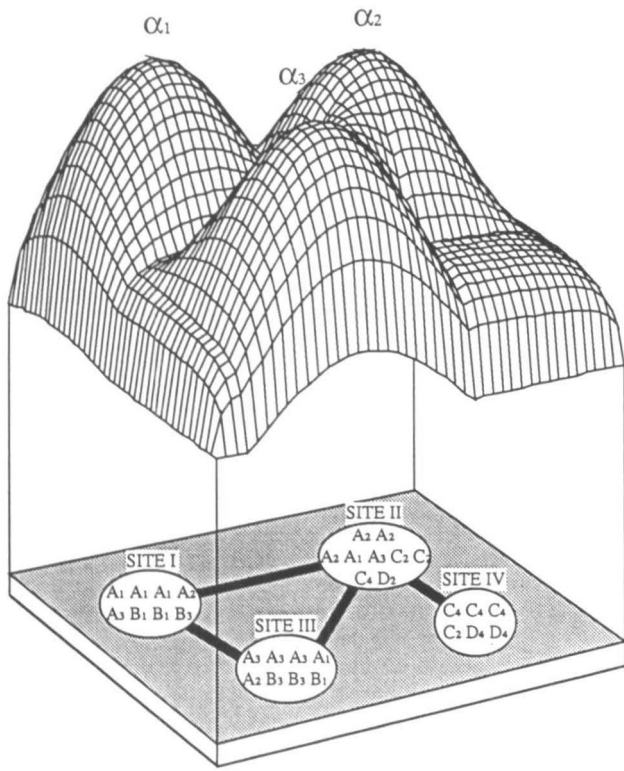


Figure 1 (Mesulam). Schematic illustration of the interrelationship between behavioral and anatomical components (from Mesulam 1990).

in site I (e.g., the encoding of exploratory movements is done mostly in the FEF but to a lesser extent also in area PG). Each site belongs to several intersecting networks (Mesulam 1990). For example, function C is distributed in an intersecting network that includes site II. The behavioral (or cognitive) components of alpha are designated as α_1 , α_2 , and α_3 and may correspond to exploratory behavior, perceptual representation, and motivational mapping, respectively. The peak of α_1 is approximately over site I, but there is also a skirt that extends into the other two sites. The resulting topological plane (with peaks α_1 , α_2 , and α_3) corresponds to the clinically observed

behavior. Recent anatomical observations indicate that the interconnections among the relevant anatomical sites of a large-scale network are organized in a manner that can sustain parallel distributed processing (Morecraft et al. 1993). These interconnections and the process of diaschisis indicate that a perturbation in one anatomical component will tend to be propagated (along a spatiotemporal gradient) to the other components of a network. The vertical organization of the anatomical, computational, and cognitive planes is depicted in Figure 2. It is important to point out that Farah's "modules" reside mostly within planes 2 and 3 whereas my "network components" are located predominantly at the anatomical plane.

Figures 1 and 2 suggest that the anatomical mapping of behavior is both localized and distributed but neither equipotential (holistic) nor modular (insular or phrenological). According to Figures 1 and 2, each site in association cortex belongs to several intersecting networks so that an individual lesion, even when confined to a single cytoarchitectonic field, is likely to yield deficits in multiple domains. Posterior parietal lesions, for example, cause deficits in complex visuospatial processing in addition to unilateral neglect. Conversely, some lesions (or electrical stimulations) may remain behaviorally silent under certain conditions because alternative parallel channels may become available. The model in Figures 1 and 2 helps to explain how anatomical localization is compatible with the fact that lesions in different parts of the brain can yield perturbations of the same overall behavioral domain, why single lesions lead to only partial deficits of a given behavior or to multiple behavioral deficits, and why functional mapping studies are likely to detect multiple areas of activation even when the subject is engaged in a single task. For the more practical purposes of neuropsychological assessment, this model predicts that no neuropsychological task (or neurocognitive paradigm) can ever be entirely specific for a single region of association cortex and that the clinician need not look for multiple lesions just because the patient shows more than one cognitive deficit.

This model does not deny the existence of local specializations in the brain but it implies that individual behaviors are represented in multiple sites and that each site subserves multiple behaviors, leading to a distributed and interactive but also coarse and degenerate (one-to-many and many-to-one) mapping of anatomical substrate onto neural computation and computation onto behavior. This distributed mapping may provide an advantage for the rapid computation of complex cognitive operations and sets the network approach sharply apart from theories

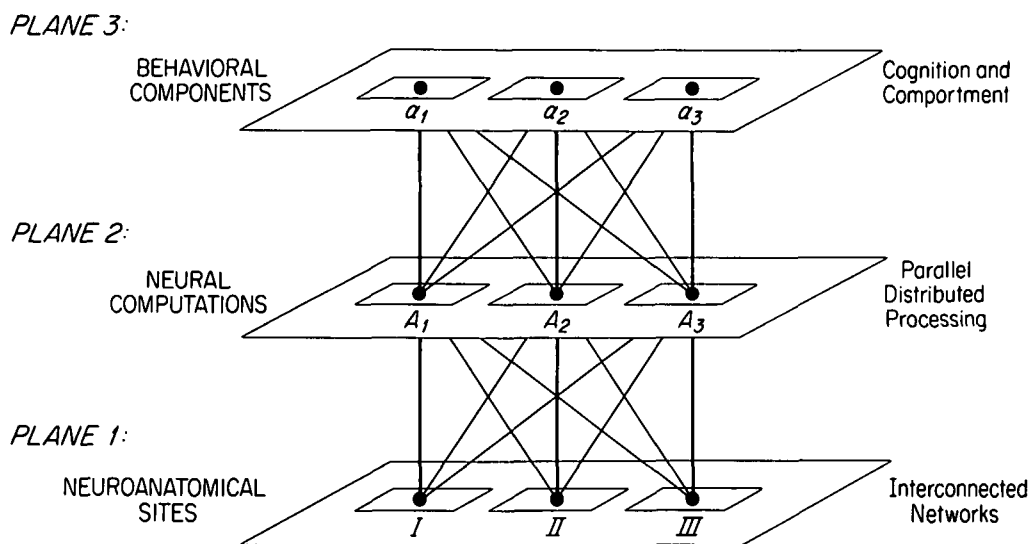


Figure 2 (Mesulam). Schematic illustration of interrelationships among anatomical, computational, and behavioral planes (from Mesulam 1990).

that postulate a one-to-one relationship between behavior and anatomical site. The neural connectivity among network components, the intersection of networks, and the physiological process of diaschisis clearly show that the version of the locality assumption described in the target article is not compatible with what we know about the organization of the real brain.

ACKNOWLEDGMENTS

I want to thank Leah Christie for expert secretarial assistance. This work was supported in part by grant NS30863 from the NIH.

Computational levels again

Mike Oaksford

Cognitive Neurocomputation Unit, University of Wales at Bangor, Gwynedd LL57 2DG, United Kingdom

Electronic mail: pss027@bangor.ac.uk or mike@cogsci.ed.ac.uk

My purpose in this commentary is to argue that Farah's critique of the locality assumption unnecessarily represents "box-and-arrow" accounts of cognitive processes as being in conflict with recent connectionist models of neuropsychological phenomena. I argue that behind this putative dispute is a conflation of levels of analysis of complex computational systems such as the human brain. Although accounts of such levels are contentious, I will stick with Marr's tripartite division as the best known, most influential, and most clearly stated. Briefly, what functions need to be computed in the performance of some task are specified at the *computational* level; how those functions are to be computed is described at the *algorithmic* level (this level also includes the specification of the representations over which the algorithms are defined); and how the algorithms are physically realised is outlined at the *implementational* level. As emphasised by Marr (1982, and, e.g., Chater & Oaksford 1990; Oaksford & Chater 1991; Rumelhart & McClelland 1985), but *pace*, for example, Fodor and Pylyshyn (1988), these levels are not autonomous, but constrain each other so that a complete psychological explanation of any task performance must invoke all three levels of analysis. I will now attempt to argue that the dispute Farah tries to resolve in favour of connectionist systems need not arise and is based on a failure to appreciate the importance of all three levels to psychological explanation. To do so I concentrate on the first model she discusses, which accounts for the dissociation between the knowledge of living and nonliving things.

Farah shows how a parallel distributed processing (PDP) model with separate modules for visual and functional semantics may account for the living/nonliving dissociation *and* for the fact that patients with a visual knowledge deficit for living things also show a deficit in functional knowledge of living things. In accounting for the latter observation the network is only lesioned in the visual semantics module, demonstrating that a single lesion in one part of the model may lead to suppressed performance in another undamaged part. This is important, because, in direct contradiction of locality, it demonstrates that "the remaining unimpaired processes will work differently when one component is not functioning normally."

This argument does not refute locality, however, if locality is defined at the computational level and not at the implementational level. According to locality, the observed behaviour involves a normal model that is subject to the effects of the lesion. Lesion effects are mapped out at the computational level by the pattern of dissociations a patient reveals. So if a patient is impaired on a task that requires function A but not on tasks requiring functions B and C it can be assumed that the lesion has affected A but not B or C, in which case a cross can be placed on the box representing function A. The evidence hence implies that a cross should be drawn through the visual semantics box. To account for the deficit in functional knowledge of living things, it must be assumed that functional semantics *fractio-*

notes into knowledge of living and nonliving things. Hence a cross is drawn through the sub-box in functional semantics for living things. This seems to be the local account to which Farah objects.

And indeed it does seem unparsimonious. If nothing else, it reestablishes appeal to the specific living/nonliving distinction that the more general visual/functional distinction was intended to avoid. However, it does appropriately summarise the patients' pattern of deficit at the computational level. And, of course, the computational level does not specify how a function is to be implemented. (Note that because a task is decomposed as requiring two functions to be computed, this does not mean that, ipso facto, two anatomically distinct causal mechanisms are required for their computation.) In particular, the local computational level description provided above is consistent with Farah's nonlocal implementation. This is because all the local computational level model states is that functional knowledge of living things is impaired but functional knowledge of nonliving things is not, which is consistent with the observed pattern of task performance.

Some confusion may arise because crosses would be placed on both the visual semantics module and the functional knowledge of living things submodule. All that this indicates, however, is that these functions are impaired, not that the physical mechanisms upon which these functions supervene have sustained physical damage as a result of the lesion. The confusion is compounded by the problem that although locality is best treated at the computational level, Farah uses it at the implementational level, adducing evidence from "the highly interactive nature of *brain* organization" to refute it. Thus it seems that two notions of locality are being discussed: computational locality and implementational locality. The problem is that nonlocal PDP implementations are consistent with the retention of computational locality. Computational locality allows neuropsychological evidence to bear on the functional decomposition of the tasks that have been examined with normals. It is intended to rule out the possibility that damage results in a wholly new *functional* organisation – a possibility that would indeed invalidate any inferences from neuropsychological data to normal, computational level models.

In addition, Farah's models are themselves local at the computational level. She argues that lesions to the visual semantics module alter the function of the functional semantics module. Qua computational level description that seems to be simply false. Whereas the inputs to the module have changed, the actual function it computes remains the same. To establish that the function the module computes has changed would require some attempt to define the functional equivalence of two systems. An obvious first stab is that two systems compute the same function if and only if, given the same inputs, they produce the same outputs (at least this is adequate at the computational level where it is only the functions *in extension* that need to be specified – *algorithmic* equivalence is an altogether more complex issue (see, e.g., Foster 1992). For Farah to establish that the functional semantics module computes a different function she would have to show that it produces different outputs after the lesion to the visual semantics module than it did before the lesion to the visual semantics module, *given the same inputs*. Since the lesion to the visual semantics module simply alters the range of inputs to the functional semantics module, it must compute the same function before and after the lesion. In sum, functionally, that is, at the computational level, Farah's models are local, although at the implementational level they are distributed and interactive. In this respect it is important to note that the term "locality" invites confusion because of the association with local versus distributed representation. However, this representational point is not at issue here. Farah's models are not "nonlocal," in her sense of the term, simply by virtue of using distributed representations.

In conclusion, if classical "box-and-arrow" diagrams and PDP

models are viewed as descriptive of the same processes but at different levels of analysis of a complex system, the dispute Farah identifies dissolves. This is not to say that such an account would meet with universal acceptance. There is a great temptation to reify box-and-arrow diagrams as claims about actual causal structure. However, carefully distinguishing between these levels of analysis and recognising the necessity of each may prevent another protracted and ultimately futile internal squabble. Complete psychological explanations of cognitive phenomena require explanations at each of Marr's three levels (Chater & Oaksford 1990; Rumelhart & McClelland 1985). Hence, local box-and-arrow computational level descriptions need not be viewed as theoretical competitors to PDP implementations.

Parallel distributed processing challenges the strong modularity hypothesis, not the locality assumption

David C. Plaut

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213-3890

Electronic mail: plaut@cmu.edu

Farah presents convincing arguments that principles of parallel distributed processing (PDP) provide more parsimonious explanations of a number of neuropsychological phenomena than do traditional modular accounts. She ascribes this to the fact that PDP systems violate the "locality assumption" in that damage in an interactive network can have nonlocal effects. On closer inspection, however, Farah has misinterpreted the locality assumption as it has been formulated in the literature. Furthermore, Farah's version of the locality assumption is violated by modular systems as well, and thus does not provide a useful basis for distinguishing PDP and traditional accounts. Rather, this contrast is better understood at a more general level, in terms of a rejection of the strong modularity hypothesis (e.g., Fodor 1983). In particular, the most fundamental contribution of PDP modeling to neuropsychology is that it allows a principled expression of nonmodular, interactive computation in which the analysis of the effects of damage is tractable.

According to the strong modularity hypothesis, the cognitive system is composed of informationally encapsulated components that receive input from only a few other components and produce all-or-none output in discrete stages. Informational encapsulation entails that the knowledge required for a process is available only to the component dedicated to that process and that any partial results of the process are unavailable to other components. Each of the central properties of PDP systems that Farah lists (sect. 1.4, para. 2) contrasts in a specific way with these properties of modular systems: (1) the knowledge involved in a process is *distributed* in connection weights throughout the network rather than being localized to a particular component; (2) processing is *graded* and continuous rather than being staged and all-or-none, making partial results in one part of the network continually available to other parts; and (3) groups of units representing different types of information are highly *interactive*, instead of receiving inputs from only a few other components. Furthermore, while the strong modularity hypothesis says little about the nature of processing within each component, PDP systems use a common set of computational principles both within and between groups of units: processing takes the form of graded interactions among distributed patterns of activity.

In arguing that PDP systems provide better accounts of neuropsychological data than modular systems, Farah focuses on what she calls the "locality assumption," which she interprets as implying that the effects of damage are local; that "nondamaged components will continue to function normally" (sect. 1.1, para. 1). This statement can be interpreted in two ways: (1)

nondamaged components *behave* normally, in that their output to other components is unchanged, or (2) nondamaged components *compute* normally, in that their input/output function is unchanged. Farah clearly has interpretation (1) in mind. The central claim of Farah's target article is that the advantage of PDP accounts stems specifically from the occurrence of nonlocal effects of damage within these systems, in violation of the locality assumption. Yet clearly the nondamaged portions of the networks compute normally but behave abnormally in response to corrupted input from damaged portions.

The same is true, however, of nondamaged components which receive input, either directly or indirectly, from a damaged component in a modular system. For example, in de Haan et al.'s (1992) model of face processing (see Fig. 10, sect. 2.3.2, para. 1), no one would claim that the "response effector systems" would continue to function normally if, say, "structural encoding" were impaired. In fact, Farah acknowledges this point, stating that the effects of damage in such a system are confined to "lesioned components and the relatively small number of components downstream" (sect. 1.3, para. 1, emphasis added). That damage alters the behavior of intact components downstream clearly violates Farah's interpretation of the locality assumption. But note that the claim that there are relatively few such components does not come from the locality assumption per se but from more general assumptions about modular systems: processing tends to be feedforward (staged) and each component receives few inputs. In this regard, PDP systems differ from modular ones only in a quantitative way: portions of a PDP network typically receive a wider range of input than do components in a modular architecture. In both frameworks, however, nondamaged processes will be affected by corrupted input from damaged processes, and thus PDP and modular systems are on equal footing with respect to Farah's interpretation of the locality assumption.

In fact, in the neuropsychological literature, what corresponds most closely to Farah's interpretation of the locality assumption (1 above) is Caramazza's (1986) "transparency assumption," according to which it must be possible to relate the effects of damage on the *behavior* of the cognitive system to its normal operation in a principled way. This relationship had to be "transparent" in earlier formulations (Caramazza 1986), but merely "tractable within the proposed theoretical frameworks" in later ones (Caramazza 1992, p. 82). The locality assumption was originally formulated in the context of the transparency assumption, where it corresponds most closely to interpretation (2) above – that the damage itself must be local.

My formulation of the transparency assumption implies that [neuropsychological evidence] E_i can only be related to [a cognitive model] M when the damage to the system is "local." This assumption may be too strong as an *in principle* claim – nonlocal, very general modifications of the system may still allow the possibility of relating E_i to M . However, *in practice*, given the tremendous complexity of the systems we are dealing with, it may *only* be possible to draw meaningful conclusions from impaired performance to normal cognitive systems under a restricted sort of condition. (Caramazza 1986, p. 52, emphasis in the original)

Thus, the standard locality assumption is simply a way to ensure that the transparency assumption is tractable, and the transparency assumption is simply a way to ensure that the effects of damage are interpretable. In this light, PDP modeling in neuropsychology is important, not because it is "not constrained by the locality assumption" (sect. 1.3, para. 5) as Farah contends, but rather because it provides a rich, nonmodular theoretical framework in which it is nonetheless possible to relate normal and impaired behavior in a principled way.

Critically, all the advantages of Farah's PDP accounts can be most naturally understood as arising from ways in which the nature of computation in these systems violates various aspects of the strong modularity hypothesis. In the simulation of both semantic memory impairments (Farah & McClelland 1991) and

impaired attentional allocation (Cohen et al., in press), the ability of the networks to account for the data arises out of graded cooperative and competitive interactions among portions of a network that are not meaningfully interpretable as informationally encapsulated components. Furthermore, Farah acknowledges that the success of the simulation of impaired face processing (Farah et al. 1993) stems less clearly from violating her interpretation of the locality assumption. On the other hand, the fact that residual knowledge after partial damage can support performance on implicit tasks stems directly from violating a central aspect of the strong modularity hypothesis: knowledge is not encapsulated in separate components but distributed throughout the network (see Hinton & Shallice 1991; Plaut & Shallice 1993; in press, for similar results).

Viewing PDP systems as challenging the strong modularity hypothesis rather than Farah's locality assumption also provides a better understanding of the finer-grained analyses she mentions. In Hinton and Shallice's (1991) simulation of deep dyslexia, visual and semantic errors cooccur because the knowledge of how visual representations relate to semantic representations is distributed throughout the network rather than being confined to a "visual" component and a "semantic" component, respectively (see Plaut & Shallice, in press, for further results and discussion). In Patterson et al.'s (1989) simulation of surface dyslexia, poor performance on low-frequency exception words and the occurrence of regularization errors occur because the knowledge of all spelling-sound correspondences is embedded in the same set of weights rather than being split into "lexical" and "nonlexical" components, and the robustness of a given correspondence in the face of damage depends on its frequency of occurrence (but see Behrmann & Bub 1992 for criticism of the adequacy of the account). In Mozer and Behrmann's (1990) simulation of neglect dyslexia, the lexicality effects after visual damage arise simply from the fact that lexical knowledge can reconstruct corrupted word input but not corrupted nonword input. The only sense in which damage has nonlocal effects in any of these simulations is the same one that applies to modular systems: intact components downstream from a lesion are affected by corruption of their input.

In summary, I strongly agree with Farah that PDP principles provide a way of characterizing cognitive processes that is fundamentally different from more traditional, modular frameworks, and that systems which embody these principles can generate more satisfactory accounts of a wide range of psychological and neuropsychological phenomena. However, I disagree with her about the specific aspects of PDP systems which distinguish them from modular systems. Both PDP and modular systems exhibit nonlocal effects of damage and thus violate Farah's interpretation of the locality assumption. However, the distributed, interactive, graded processing in PDP systems contrasts sharply with the encapsulated, staged, all-or-none processing in systems adhering to the strong modularity hypothesis. It is exactly these distinctions, and not a violation of Farah's locality assumption, that are fundamental to the strengths of the PDP approach in cognitive neuropsychology.

ACKNOWLEDGMENTS

I would like to thank Marlene Behrmann, Jay McClelland, and Tim Shallice for helpful comments on an earlier draft of this commentary.

Local and distributed processes in attentional orienting

Michael I. Posner

Psychology Department, University of Oregon, Eugene, OR 97403
Electronic mail: mposner@oregon.uoregon.edu

Farah argues that lesion data have usually been interpreted according to two principles of localization: (1) the effects of the

brain damage are local and (2) and nondamaged parts of the system operate normally. To show how studies of brain injury and neuroimaging combine to provide a different form of understanding of localization, I would like to consider the finding that parietal damage affects attention to visual locations.

Farah's target article provides an alternative explanation for my suggestion that the ability to disengage attention is damaged by a lesion of the parietal lobe (Posner 1988). When we first made this suggestion it was not at all clear that there would be a local mechanism related to the task Farah describes. Although lesions of the parietal lobe did influence the ability to disengage, it was certainly possible, even likely, that this was due to a nonlocal mechanism. Indeed, reaction time data from patients with lesions of the parietal lobe, thalamus, and midbrain showed that each of these lesions had some influence on the task, although they did not appear to involve the "disengage" operation (Posner 1988). Our description from the start was in terms of an interactive network in which parietal mechanisms were in close coordination with thalamic and midbrain areas.

Recently the idea of localized mechanisms in the parietal lobe has been supported by PET (positron emission tomography) studies using variants of the model task first used for patients (Corbetta et al. 1993). It seems clear that when normal subjects shift attention from one location to another, whether to report targets or merely because they are free to process transient information, there are strong increases in blood flow within a restricted area of the superior parietal lobe. For left visual attention shifts only the right parietal lobe is active whereas for right visual field attention shifts both left and right hemispheres are active. This is strong evidence for a very local effect in normal subjects. Perhaps this could be the "attention" box in the Farah architecture. Yet this brain area was not active during many other tasks that involved attention to color, form, motion, or location but did not require a shift of attention from one position to another. The PET data suggest that the parietal lobe is crucial when subjects disengage from one location to move to another. They do not separate the disengage operation from the move operation, although Robinson et al. (1991) has recorded from parietal cells that seemed to support an attentional operation related to disengagement from a receptive field location.

Does this mechanism require connections from a location in the opposite visual field, as suggested by the connectionist model Farah presents? It is noteworthy that Luck et al. (1989) have shown that when the corpus callosum is severed, the search for a conjunction target occurs within each visual field. There is reason to believe that conjunction search involves covert shifts of attention between locations, so this result implies that the mechanisms of shifting attention do not depend on connections across the midline, as implied by the Farah model. However, the unified visual-orienting system of the normal subject is split by the lesion. Our initial experiments involved two horizontal locations, one in each field, but we were able to show that parietal damage affects all locations in both visual fields on those occasions when the target is in a contralesional direction relative to the current focus of attention. If each position in a visual field can serve as a cued location, Farah's model would have to involve not a few connections but many hundreds of them. Would it scale up to "handle" this problem, and if so, would it provide any fresh insight? These are the kinds of questions to ask in determining whether one would prefer a connectionist model to evidence for a local mechanism coming from a convergence of research between imaging and brain damage studies.

A further factor shows the great advantage of connectionist networks. Even if there is a local occurrence of a disengage operation, it is important to understand the communication between that operation and directly related ones. A disengage function would not be useful unless it could influence information accumulating at quite distant sites about the exact form of the stimulus to which attention is switched. According to La-

Berge (1990), this involves connections from the parietal system to the inferior temporal cortex via the pulvinar nucleus of the thalamus. The expectations that lead to attention to a location operate in a distributed network. It appears, however, that the connections are between areas performing identifiable sub-routines or operations. We must understand both the local computations and their coordination in real time to understand the network functions. Noninvasive electrical and magnetic recording, if coordinated with PET or other spatial methodologies, allow us to study the temporal order of anatomically limited computations (Posner & Rothbart, in press).

To understand brain injury we must understand that the reduction of an operation at one part of the network can have important consequences for other parts of that network. For this reason no extreme form of modularity that postulates encapsulated vertical networks passively activated by input will do. If we only had data from brain damage and no way to image the normal brain we would have great difficulty in determining what was the effect on the local computation and what was the remote effect of the damage. A combined approach to normal and pathological functions instead relies on anatomical and circuit tracing methods to supplement patient studies.

Perception and its interactive substrate: Psychophysical linking hypotheses and psychophysical methods

Robert Sekuler

*Department of Psychology, Brandeis University, Waltham, MA 02254;
Department of Biomedical Engineering, Boston University, Boston, MA 02215*

Electronic mail: sekuler@binah.cc.brandeis.edu

Farah's architectural criticism put me in mind of John Donne's sermon more than three and a half centuries ago: "No man is an island, entire of itself; every man is a piece of the continent, a part of the main; if a clod be washed away by the sea, Europe is the less, as well as if a promontory were. . . ." With a little imagination one could construe Donne as preendorsing Farah's cautionary message about a highly interconnected and interactive brain.

Farah began by arguing that no part of the brain is an island, entire of itself. Worried that this fact might spell the demise of neuropsychology, Farah saved the day (and perhaps the field) by showing that the inconvenient complexities of anatomy and physiology can be dealt with in a principled and enlightening manner. It is certain that neuropsychologists will weather these inconveniences. But what about other researchers? For example, what about the various species of psychologists who study normal rather than brain-damaged subjects? Should they assume that the problem is someone else's and thus disregard Farah's paper? The answer is "most certainly not." In my view, Farah's thesis is crucial to much of contemporary psychology.

To one degree or other, many of psychology's subdomains are concerned with possible connections between psychological phenomena and the body. For some subdomains, including perception, theorizing about mind-body connections occupies the intellectual foreground; arguably, it's where the action is now. For other subdomains, including personality, the real impact of such theorizing waits for a moment in the future; even then, of course, it may never be front and center.

Though the timetables may differ, sooner or later researchers in most areas of psychology will have to confront the interactive brain. This confrontation will be difficult. After all, psychology now assumes a brain that is barely recognizable as the interactive one Farah urges us to embrace. If Farah's admonitions hold for fields outside neuropsychology, and I think they do, much of psychology plainly needs straightening out. To see what I mean,

consider some implications for the study of perception by normal, nonimpaired observers.

Perceptionists take great pride in propositions that link perceptual and physiological states. These propositions have played a central role in guiding the field's development. In the modern era, psychophysical linking propositions can be traced back one hundred years to G. E. Müller. Later, these propositions were operationalized by Boring (1942) and systematized by Teller (1984). Although one can imagine linking propositions that recognize the force of Farah's argument, all linking propositions currently in use reflect two assumptions: first, that traffic in the nervous system is one-way, and second, that it is possible to make a causal link between a perceptual state and activity at one neural locus, or at most a few neural loci. If these assumptions were examined and abandoned, the resulting linking propositions would be dramatically different from those currently in use.

Much of perception ignores the massive evidence for rampant parallel and reentrant pathways that characterize every mammalian sensory system. For example, even the briefest glance at an up-to-date wiring diagram of the mammalian visual system (Felleman & van Essen 1991) is enough to convey the complexities inherent in the hundreds of interconnections that link the system's many sites. To date, though, there have been few systematic attempts to prospect for the perceptual interdependencies entailed by the reinforcing webs and overlapping tangles of the wiring diagram. The reasons for this theoretical immaturity are undoubtedly heterogeneous, but they may well include the undeniable fact that the simple view of the brain makes for a better story, a narrative that is easier to express and certainly easier to comprehend (Kelley 1992).

If backward notions about psychophysical links have retarded perception's advance, psychophysical methodology also deserves some of the blame. Here I have two particular issues in mind. One likely consequence of the mind's normal functional architecture is that any stimulus, no matter how simple, brief, or impoverished, generates a number of distinctly different effects in the nervous system. These multiple effects develop at different rates, yield different end products and, as a result of the system's wiring, may also be differentially susceptible to influence by the observer's expectations, by training, and by other nonsensory variables.

In her discussion of prosopagnosia, Farah showed that seriously impaired face recognition can coexist in the same person with near-normal face recognition. Which side of the coin one sees depends upon whether face recognition is assayed via explicit or implicit measures (sect. 2.3). In most fields of psychology, an array of diverse measures is essential to capture the gamut of diverse effects that are initiated by any single stimulus.

But psychophysics, visual, auditory, or otherwise, is particularly handicapped in this regard. Standard psychophysical methods are designed to quantify conscious experience – that's what Fechner was aiming for. All the classical methods depend upon the observer's conscious experience and equally conscious decision. Some techniques can open a window onto nonconscious perceptual processing (Kunst-Wilson & Zajonc 1980; Mandler et al. 1987), but for most students of perception these results are off the intellectual radar screen, out in the field's periphery. Although these studies are intriguing, their outcomes are treated as exceptions or oddities rather than as signs of influences that are probably ubiquitous, though hard to measure. To paraphrase John Donne (slightly incorrectly), it seems to me that psychologists, including perceptionists, should not have to ask for whom Farah's theoretical bell tolls: it tolls for all of us.

Locus-pocus (which and whose locality assumption?)

Carlo Semenza

Dipartimento di Psicologia Generale, Università di Padova, 35139 Padova, Italy

Electronic mail: psico28@ipdunivx.unipd.it

It may be an easy prediction that a direct comparison between modular and connectionist accounts of neuropsychological dissociations will become commonplace in forthcoming years. Farah's opening of a forum on this subject is therefore welcome and of major interest. Her target article, however, would have been more effective, and, probably, the position she takes more appealing, if not centered on a critique of what she calls the "locality assumption." The way this matter is presented misleads the reader throughout Farah's discussion because, at present, nobody would subscribe to the locality assumption as Farah seems to intend it. Thus, although she can readily criticize the locality position, her whole argument is unduly undermined.

As formulated by Farah (without citing anyone else), the locality assumption states that the effects of brain damage on the functional architecture are local, that is, the nondamaged components of the architecture continue to function as they did before the damage. According to Farah, such an assumption would be justified in terms of informational encapsulation, the key attribute of modularity. Indeed, the overall message of Farah's paper is that since in a few neuropsychological deficits the locality assumption does not seem to lead to the more parsimonious explanation, and should therefore be abandoned, the modularity assumption, from which the locality assumption follows, should likewise be rejected. This inference is fallacious: I assume that several commentators will pick this up so I will not elaborate further on the subject; nor will I consider the merit of alternative explanations for the specific cases Farah has discussed. What I would like to point out, instead, is that the locality assumption and the closely related transparency assumption (also misrepresented in Farah's paper) have been formulated otherwise. In particular, the inference of the form "selective deficit in ability *A* implies a component of the functional architecture dedicated to *A*" has not merely been cautioned against but rejected as logically untenable, because it is valid only under certain conditions.

Semenza et al. (1988) specifically argued, for example, that the identification of a deficit (effect) *D* does not imply that any separate processing component *P* exists. *P* may merely correspond to a quite loose definition of processes that are supposed to be part of normal performance. Yet, for the observed effects there may be a cluster of alternative architectural solutions and impairments; hence if *D* occurs selectively, it may be viewed as an independent phenomenon, but this in no way entails that *P* is an independent process. The interpretative framework of the locality assumption in Farah's formulation is founded on a confusion between the dissociation of phenomena (which can actually be observed), the dissociation of processes (which are generally unspecified), and the definition of inherent operations (which, in fact, is based on an a priori conception of a given task). This framework is based on an illusion of transparency that stems from arbitrarily mapping a priori concepts onto an interpretative situation.

The assumption of transparency was indeed ambiguously formulated at the very beginning (e.g., Caramazza 1984), in a way that could be understood as compatible with the locality assumption Farah criticizes (i.e., that there is a transparent relationship between deficit *D* and processing components *P* that are supposedly removed as a consequence of brain injury). This interpretation of the transparency assumption, however, seems to entail a paradox (Semenza 1993): what becomes transparent (thus more understandable) after a brain lesion is that which is lost! A weaker but more correct formulation may stress

the fact that brain damage allows one to better determine the nature of processes that are opaque in normals' error-free performance. The paradox would disappear: what may be transparent is what is left. Caramazza's (1992) statement that transparency requires that the behavior of the damaged system be understandable in terms of the functional architecture of the normal system is just a necessary proviso; that is, dissociations remain meaningless without a theory as opposed to just a statement of the transparency assumption itself.

Under any assumption, however, it seems unlikely that the performance of neurological patients merely represents a combination of intact and impaired patterns of behavior. In particular, contrary to Farah's beliefs, that nondamaged components of the architecture continue to function as they did before damage does not follow from the modularity assumption. Indeed, under the modularity assumption, the working of nondamaged modules may undergo considerable modification. Suppose patients have lost a module totally or partially, generally used to perform a task: they may nonetheless want to try to carry on the task (patients often do not even know there is anything wrong with them). To do so, they may use other modules that, perhaps less satisfactorily, allow the task to be completed. This possibility makes things more complicated, but it may help research. The functioning of residual modules could even be highlighted by brain damage; overworking may somehow make them more transparent to the observer (Marin et al. 1976; Saffran 1982; Semenza et al. 1988). When this is not the case, it is true that the researcher's life is harder, because a number of methodological heuristics (see Shallice 1991) must be employed to make the process used by the patient more transparent. The task is not impossible, however.

In conclusion, I have tried to present a fair account of the locality assumption, one that may correspond more closely to people's thinking than the version presented by Farah. I think that this account is much less prone to her criticisms. I assume that for the specific examples she gives, where, on her account, modularity seems to lead to less preferable explanations, she will be directly answered by the authors of the competing theory, who, I suspect, are still, like me, awaiting better reasons (or magic) for believing one theory or the other.

Throwing out the neuropsychological data with the locality bathwater?

Philip Servos^a and Elizabeth M. Olds^b

^aDepartment of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2 and ^bDepartment of Psychology, Stanford University, Stanford, CA 94305-2130

Electronic mail: pservos@cogsci.uwo.ca; olds@psych.stanford.edu

Farah's point is well taken that one should be aware of the possible pitfalls of the locality assumption in neuropsychological research. Her parallel distributed processing (PDP) models provide elegant, distributed alternatives to various hypotheses that depend on the locality assumption. Although her basic arguments are sound, the extent to which Farah objects to any form of the locality assumption remains unresolved.

One problem with Farah's arguments is that she seems to address only a very extreme version of the locality assumption. Surely few neuropsychologists would claim that a given structure *wholly* and *independently* subserves a given function, especially the reasonably high-level functions Farah discusses. The incredibly rich interconnectivity of the brain would be enough to convince anyone otherwise. Clearly, a given structure would project to and receive inputs from various brain regions, both proximal and distal, which might well play important roles in the function under scrutiny. Most of us would accordingly be content with the claim that a given damaged structure is *import-*

tant, or *critical*, for a given disrupted process. If a functional deficit frequently (or always) arises following damage to one brain region, and *not* when damage occurs to other brain regions, it seems reasonable to assign a critical role to that brain region in the processes underlying that function. We can make progress in this way even though we cannot be absolutely sure there is a one-to-one mapping between structure and function. This weaker version of the locality assumption seems to be the one most neuropsychologists would adopt (Farah's claims to the contrary notwithstanding). It seems surprising, then, that even though Farah outlines several weaker versions of the locality assumption early in the target article, she proceeds to address only the extreme version of it.

Another problem with Farah's arguments against the locality assumption is that, in the extreme, they preclude much if not all neuropsychological research. Farah seems to be making the strong claim that because it can be difficult to ascertain whether a given brain region is wholly responsible for a given function, we should avoid using the locality assumption and hence any approach that makes use of it. We agree that a strict modularity of function is probably incorrect and that cognition in general is more distributed and interactive than some might claim. However, that does not mean that a given cognitive process involves such diffuse and nonlocalized processing that it is futile to use neuropsychological techniques to identify the brain structures underlying it. Furthermore, we are not sure what would replace the locality doctrine, given that most corroborative techniques (e.g., neurophysiology and functional neuroimaging) also make use of this framework. At present, PDP and other modeling techniques do not yield solutions that demonstrate the specific neuroanatomical loci for particular brain processes, although they have clearly shown potential as powerful tools for computationally testing theories of human cognition; they may also be able to *guide* neuropsychological research by providing models of the underlying processes (e.g., see Gluck & Rumelhart 1990). Obviously, the plausibility and relevance of such models are increased when they incorporate details about the neural architecture ultimately responsible for the processes in question. Farah is clearly not denying the importance of neural constraints in modeling brain function but it is unclear how she would like to see this sort of data collected, if not by techniques that make use of the locality assumption.

Although Farah argues against a hard version of the locality assumption and is quite sympathetic to the distributed framework, it is unclear to what extent she embraces the latter perspective. Is she, for example, willing to entertain an extreme version of the distributed framework – something akin to Lashley's notion of equipotentiality? We suspect not. Both extreme positions seem untenable. This then places her in a position that is not unlike some weaker version of the locality assumption that most neuropsychologists would hold anyway. Moreover, is she contending that the extent of interactivity is constant across all kinds of cognitive processes, or just for the kinds of examples she chose – examples for which locality-based explanations may be particularly inappropriate?

In addition, though Farah is quick to reject the concept of modularity, she still seems to accept it implicitly to some extent. For example, in one passage she states that "it is well known that different brain areas are dedicated to representing information from specific sensory and motor channels." So, at some level, at least for what might be called lower functions, she implies that there may be some functional modularity in the brain. She also seems to make use of modularity in her own work in that each model she develops accomplishes a specific task. That is, although the effects of damage to a given system are less local than some neuropsychologists might feel comfortable with, they are still local to the system ("module"?) she is modeling. This suggests that her distributive logic may not apply equally well to all levels of brain organization; that is, at some level, modularity may exist. Thus, although Farah raises an extremely important

issue in cognitive neuroscience – the problem of establishing which brain regions are considered essential for a given cognitive process and which are irrelevant – her models do not provide clear answers (although, to be sure, this is not a trivial task). She opposes the idea of processes being encapsulated and instead believes in the interactivity of structures, but it is unclear to what degree she holds that such interactivity exists.

We are certainly not arguing here for modularity or any extreme form of the locality assumption. Farah's arguments to the contrary are quite strong and elegant in this regard. However, there needs to be further clarification of what locality represents and when and to what extent it is relevant in the discourse of structure/function relationships. Discounting neuropsychological data collection just because some particularly extreme form of a locality assumption is not feasible seems a little like throwing out the baby with the bathwater. Instead, neuropsychological research should obviously be continued, but we should consider deeply the possibility that many mechanisms may be more distributed than we thought, in the ways that Farah and others have described. That is, there may be a reasonable compromise: we should probably continue to use some form of locality assumption to guide empirical research, along with techniques that stress nonlocalist assumptions to interpret the data (e.g., Rumelhart & McClelland 1986). Farah's arguments are quite compelling, in that they push neuropsychologists toward the "distributed" as opposed to the "local" end of the interpretive continuum of cognitive functions. What remains to be discovered is where exactly on this continuum we should set our feet.

The real functional architecture is gray, wet and slippery

Steven L. Small

Cognitive Modelling Laboratory, Department of Neurology, University of Pittsburgh, Pittsburgh, PA 15261-2003

Electronic mail: sls@cs.cmu.edu

The target article by Farah argues against the view of the mind as a set of "relatively independent information-processing subsystems" and the assumption of this view in the design of neuropsychological studies and the interpretation of their results. I am sympathetic to Farah's conclusion, as are many neurologists, speech and language pathologists, and computer scientists. However, a number of her arguments are weakened by missing neurobiological and computational links.

1. Anatomical localization. Neurologists and psychologists have periodically attempted to explain the relationship between local brain processes and global ones. Advocates of strong anatomical localization of function (Benson & Geschwind 1989; Broca 1861) produced much data from a clinico-anatomical perspective, attempting to correlate particular cognitive behaviors (i.e., clinical syndromes) with particular anatomical areas in the human brain. This has typically involved the study of patients with naturally occurring brain lesions and has required the correlation of anatomically or physiologically described neuronal loss with descriptions of behavioral loss (Damasio & Damasio 1989; Kertesz 1983). Many neurobiologists have questioned this localization approach (Freud 1891; Jackson 1878; Lashley 1950; Marie 1906) and various locality assumptions have been the subject of extensive debate for more than a hundred years.

Over the years, descriptions of neuropsychological behavior have improved dramatically, as the three empirical studies discussed by Farah illustrate. Anatomical analysis has likewise advanced, from a reliance on postmortem brain dissection to high quality neuroimaging, including dynamic imaging of cerebral blood flow and metabolism (Raichle 1989).

A disheartening aspect of this new discussion of localization is that the anatomy and physiology of the brain have been left out. The term "locality" now refers to locations within an artificial "functional architecture," rather than the brain.

2. Brain lesions. Whereas one can only debate philosophically the effects of damage to abstract architectures, one can empirically study the effects of neurological lesions on the brain. Brain lesions come in a variety of types, however, from small areas of infarction caused by occlusions of blood vessels to infiltrating neoplasms that compress and injure adjacent brain areas to diffuse and/or focal neuronal loss. The biological effects of these different types of insults are disparate, and should influence the conclusions one draws from patients who manifest them.

Farah describes several patients with category-specific deficits, for example, and mentions their underlying neurological diseases: several patients had herpes simplex encephalitis, which typically affects parts of the temporal lobes, frontal lobes, and limbic system, causing hemorrhagic infarctions in a patchy distribution throughout these areas (Damasio & van Hoesen 1985; Price 1986). Another patient had a "head injury," and presumably had axonal shearing (Auerbach 1986) in the fasciculi of the centrum semiovale. A third had a degenerative disease thought to be relatively more localized to left temporal lobe than the rest of the brain, probably involving diffuse neuronal loss in that area (Green et al. 1990).

Does knowing this increase our understanding of the cognitive manifestations of category-specific naming impairments? I suggest that it does. By talking not only about distributed representations (as Farah does) but also their instantiation in the anatomical architecture (which Farah doesn't), one can appreciate the impact of different lesions to brain computations. What sort of computational architecture allows for category-specific impairments in the face of one of three different types of damage yet produces no such deficit in the majority of patients with any of these injuries? That challenge is not addressed in discussions of functional architectures that do not account for data on the biological architecture.

3. Computational systems and metaphors. The advent of computation has led to an ability to describe processes in a formal manner, and information processing psychology has exploited this by using computation both to describe and simulate cognitive processes. Psychological theory could thus be tested by simulating behavior in a computer model, which could serve as a formal characterization of the underlying process. An astonishing aspect of the target article is the implicit assumption that a description of an abstract model is equivalent to a description of the physical process.

The modern computer is based on one type of architecture, the "von Neumann" machine, which consists of a central processing unit (CPU), a memory, an input device, and an output device. The CPU retrieves two pieces of data from memory, performs a computation, and stores the results back into the memory. Early information processing studies described human mental computations by analogy to this sequential machine.

Later investigators, interested in the parallel processing features of human thought, hooked up (conceptually) a number of such machines, ending up with a computational framework that fits Farah's definition of a "functional architecture." Such a machine incorporates information encapsulation in modules, independent computations, and a low frequency of interaction. There was no reason a priori to understand the brain as such a collection of interacting von Neumann computers, and as Farah points out, there remains no such reason.

4. Connectionism. A crucial question is the role of the connectionist (or parallel distributed processing, PDP) models (Feldman & Ballard 1982) in providing a different framework for understanding cognitive behavior. As I am sympathetic to the goals of connectionist modelling, I will mention only two points in line with my previous discussion.

First is that connectionism (PDP) is a language for describing

theories about the mind or brain, not a theory in its own right. Thus a discussion of PDP as true or false does not make sense. Are the representations and methods of PDP useful adjuncts to other descriptive techniques in conveying theoretical notions in cognition? Do computer simulations based on these methods teach us things we could not learn (or not as easily learn) any other way? I don't view PDP as a religion, requiring faith in its "truth," but I nonetheless agree with Farah that the representations and methods of PDP are important descriptive tools and that connectionist simulation has extraordinary value to inform about cognitive and brain processes.

Second, one of the main appeals of connectionism is that its basic units of discussion are "neuron-like." Computationally, this means that rather than describing systems in terms of large bulky modules of computation, with poor ability to communicate and relatively isolated functioning (i.e., distributed von Neumann architecture), connectionism allows the description of systems in terms of small units of computation such that their pattern of intercommunication to a large extent comprises the nature of the computation. Fault tolerance is a byproduct of these designs. Such a computational paradigm makes sense neurobiologically, and connectionist models can thus make use of neuroanatomical constraints – and should.

ACKNOWLEDGMENTS

Thanks to Margie Forbes, Anthony Harris, and Gloria Hoffman for help with the preparation of this commentary. The Cognitive Modelling Laboratory is funded by the National Institute of Deafness and other Communication Disorders, under grant no. DC00054. Their support is gratefully acknowledged.

The functional architecture of visual attention may still be modular

Carlo Umiltà

Dipartimento di Psicologia Generale, Università di Padova, 35139 Padua, Italy

Electronic mail: umilta@ipdunivx.bitnet

To counter Farah's case against the existence of modules in the functional architecture of visual attention, I will make the following points: the "disengage" operation is not a particularly good example of a possible modular component of that architecture. Domain specificity is as important as informational encapsulation in the search for attentional modules. To be domain specific, attentional mechanisms must interact with perceptual mechanisms, thus producing assembled modules. A model like that of Cohen et al. (in press) could simulate egocentric neglect or allocentric neglect, but not both.

The "disengage" module. In the classical paradigm introduced by Posner (1980; Farah's Fig. 5), when the target stimulus appeared at the uncued location, attention disengaged from the cued location, moved to the location where the stimulus had just appeared, and engaged at the new location. Parietally damaged patients are very slow in responding to stimuli appearing at the uncued location on the side contralateral to the lesion. This observation has led to the notion of a selective deficit in their ability to disengage attention (assuming it can be demonstrated that the other two operations are not impaired).

This notion rests on the assumption that attention is reoriented following an invalid cue. The evidence that reorienting of attention occurs, however, is not very compelling (e.g., Henderson & Macquistan 1993). It can be argued that in normal subjects the costs incurred after an invalid cue are due to an attentional gradient that peaks at the cued location and diminishes as spatial distance from this location increases (e.g., Downing & Pinker 1985). The attentional gradient may fall off very steeply on the contralesional side in parietally damaged patients. In this way, one need not invoke a lesion of a disengage operation.

This is not to deny that mechanisms for disengaging and orienting attention exist. Perhaps these mechanisms function on the basis of competitive interactions, as modelled by Cohen et al. (in press). They are cognitively penetrable, however, and thus violate the criterion of informational encapsulation. In the example provided by Farah (Fig. 5), an abrupt event in the periphery of the visual field (the cue) automatically elicits a shift of attention. Observers also have internal control over spatial allocation of attention, however, so that they can voluntarily shift attention in accordance with a cognitive cue (e.g., Umiltà 1988). In addition, an attentional shift that would be triggered by a peripheral cue can be inhibited voluntarily (e.g., Yantis & Jonides 1990).

In conclusion, showing that the disengage deficit does not necessarily depend on a lesion of a disengage module is not relevant to Farah's argument. Perhaps, parietally damaged patients do not have a disengage deficit. In any event, whatever mechanism is lesioned, being cognitively penetrable, it cannot meet one of the main criteria for modularity.

Domain-specific forms of neglect. Evidence suggesting the existence of modular components in the functional architecture of visual attention originates from other sources. One has to use a different criterion for modularity, however: domain specificity.

A typical attentional deficit in parietally damaged patients is neglect. Dissociations between visual and acoustic neglect (e.g., De Renzi et al. 1989) and between perceptual and motor neglect (e.g., Tegner & Levander 1991) have been reported.

Within the visual modality, neglect can be confined to very specific domains. Halligan and Marshall (1991) reported evidence of left neglect for near but not far space. In peripersonal space their patient showed severe neglect on conventional tests, including the line bisection task. However, when line bisection was performed in extrapersonal space, by a pointing light or by throwing a dart, neglect was abolished or attenuated. The patient did not show personal neglect either.

Guariglia and Antonucci (1992) described the opposite dissociation in a patient who had severe personal neglect in the absence of neglect for peripersonal space (extrapersonal space, in their terminology).

Halligan and Marshall (1992) went even further, arguing that dissociations can occur within peripersonal space. Two of their patients showed a classic double dissociation for two tasks (target cancellation and line bisection) that are both performed in peripersonal space.

The patient described by Cohen and Dehaene (1991) showed neglect dyslexia only for numbers. He exhibited a spatial error pattern akin to neglect dyslexia, making most of his reading errors on the leftmost digit of any number. Apart from that, the patient had no clinical indication of spatial neglect.

Another patient described by Young et al. (1990) had severe neglect in recognizing the left half of normal faces, chimaeric faces, and half-faces presented in isolation. By contrast, he did not experience difficulty in recognizing the left side of everyday objects or of car-fronts. Furthermore, there was no evidence of neglect on item-cancellation and reading tasks.

In conclusion, if one takes domain specificity as a clue to modularity, there is enough evidence that the architecture of visual attention may indeed be modular. Sometimes attention is seriously impaired in one domain but continues to function normally in other domains.

Assembled modules. Based on domain-specific forms of neglect, one can argue that isolable mechanisms exist that are in charge of allocating attention within rather restricted perceptual domains. These attentional mechanisms must interact with the perceptual components that share the same domain.

For example, a mechanism for allocating attention within peripersonal space must be interconnected with a perceptual mechanism for representing peripersonal space. The functional architecture of the visual attention system proposed by Farah (Fig. 7) contains these highly interconnected attentional and

perceptual subcomponents. This type of architecture seems to violate the criterion of informational encapsulation.

Moscovitch and Umiltà (1990; 1991), however, have argued that modules can be assembled from a collection of more basic subcomponents and that these assembled modules do not violate the most important criteria of modularity as proposed by Fodor (1983; see also *BBS* multiple review: 8(1) 1985), that is, domain specificity, informational encapsulation, and shallow output.

A functional architecture like the one outlined in Figure 7 may just depict an assembled module, with perceptual and attentional subcomponents. This assembled module performs specialized functions, does not interact with other modules, is not influenced by semantic systems, and its output is not semantically interpreted: it is domain specific, informationally encapsulated, and delivers a shallow output.

Object-centered neglect. The so-called disengage deficit does not usually occur in isolation. It manifests itself along with neglect and/or extinction (e.g., De Renzi 1982). In fact, it might be argued that the disengage deficit is the cause of extinction and neglect (e.g., Posner et al. 1984). Patients with right-parietal damage show extinction and neglect because they cannot disengage attention from the right side of space and they direct attention to the left side (e.g., Mark et al. 1988). A model like the one proposed by Cohen et al. (in press) could easily simulate extinction and neglect if the imbalance introduced in the system is such as to render it impossible for stimuli on the lesioned side ever to win the competition.

In this way, however, neglect would always occur in a fixed coordinate system. In fact, a central concern is the coordinate system(s) in which neglect occurs. It is true that parietal patients' neglect often occurs in an egocentric space, where coordinates can be retinotopic, head-centered, or body-centered, but clinical evidence suggests that neglect also occurs in an allocentric space, the coordinates of which are object-centered.

While reading texts, neglect patients may omit the left side of individual words rather than the left side of the page (Schott et al. 1966). A similar phenomenon can be observed when patients copy a series of drawings that are horizontally aligned (Gainotti et al. 1972; Ogden 1987). They sometimes reproduce only the right side of every object in the display, including objects that are on the extreme left.

Caramazza and Hillis (1990) described a patient who was left-handed and had neglect for the right side of space after a left parietal lesion. She made reading and spelling errors only on the end part of words, irrespective of whether words were presented horizontally (i.e., the right part), vertically (i.e., the bottom part), or mirror reversed (i.e., the left part).

In a study by Driver and Halligan (1991), a neglect patient was asked to judge whether two elongated shapes, shown one above the other, were the same or different. When the principal axis of each shape was vertical, the patient neglected differences on the left. When the shapes were both rotated 45° clockwise or counterclockwise, she continued to neglect differences on the left of the objects' principal axes, even if the differences fell on the right of her egocentric axes.

I do not see how Cohen et al.'s (in press) model could possibly simulate both egocentric and allocentric neglect. The task would probably be much simpler for a model based on a modular architecture. One perceptual-attentional module would represent space egocentrically and allocate attention in egocentric space. The other perceptual-attentional module would act likewise for allocentric space. One need only assume that these two modules can be lesioned independently.

ACKNOWLEDGMENT

This paper was supported by a grant from the Consiglio Nazionale delle Ricerche (FATMA Project).

Playing Flourens to Fodor's Gall

Tim van Gelder

Department of Philosophy, Indiana University, Bloomington, IN 47405
Electronic mail: tgelder@ucs.indiana.edu

The clash between neuropsychological localizers and their distributionist opponents goes back to the dawn of neuroscience. At the turn of the nineteenth century, Gall was arguing that "the brain is composed of as many individual and independent organs as there are forces of the soul" (Spurzheim 1815, p. 272). Flourens countered that "there are not . . . different seats for the different faculties, nor for the different sensations" (Flourens 1846, p. 35). Gall and Flourens were just two sweeps of a pendulum that has continued swinging to this day. Thus, almost two centuries later, Fodor [1983] resuscitates Gall's localizationist tendencies under the name "modularity," and Farah here returns with the inevitable distributionist backlash.

Localizationists tend to assign particular higher functions or capacities to particular "organs" or "modules" in the brain; distributionists tend to see such functions as corresponding only to the operation of many brain components. Localizationists tend to see the brain's functional units as operating relatively independently of each other; distributionists tend to regard the brain's components as highly interactive. There are thus two complementary ways in which localizers tend to be localist: they localize particular functions to particular components, and they see particular components as acting "locally," that is, noninteractively. (Note, for example, how Gall insisted that to each of the forces of the soul there correspond brain organs that are both *individual* and *independent*.) Distributionists' denials are likewise complementary: functions are not local to particular brain components, and components do not act locally, but rather interactively.

Farah is conducting a classic distributionist defense against excessive localizationist tendencies. She shows how there need be no proprietary components, organs, or locations for (1) memories of living things as opposed to nonliving things, (2) a "disengage" component in visual attention, and (3) visual processing as opposed to visual awareness. She does this by explaining the relevant data by means of an architecture in which brain components are interactionist, that is, nonlocal in their effects.

In doing so, she shows how foolhardy it can be to simply assume that components are local in their effects. In fact, she presents this as her primary theoretical result: that what she calls the "locality assumption" is unreliable. Yet, surely, by the very same demonstrations she is showing just as effectively that the other locality assumption – that specific higher functions correspond to particular dedicated components – is also unreliable. The only criticism I would make of Farah's generally brilliant target article is that she emphasizes one achievement at the expense of the other.

This point can be given a sharper outline by focusing on the form of inference that Farah is undermining. She claims that "the locality assumption licenses quite direct inferences from the manifest behavioral deficit to the identity of the underlying damaged cognitive component, of the form 'selective deficit in ability A implies a component of the functional architecture dedicated to A'" (sect. 1.1). If we lay out this inference explicitly, it looks like this:

Premise 1. There is a selective deficit in ability A.

Premise 2 (Farah's locality assumption, LA). The effect of damage on a component of the functional architecture is local; that is, "nondamaged components continue to function normally" (sect. 1.1).

Conclusion. There is a component of the functional architecture dedicated to ability A.

The main thrust of the target article is that inferences of this form are unreliable, and Farah lays the blame at the feet of the LA. Neuropsychologists reach poor conclusions when they rely

on the LA; it is often a better hypothesis that brain components are basically interactive.

Now, the point I am making is that there are really two locality assumptions. Farah isolates one as the LA, makes it explicit in the above form of inference, and successfully undermines it. The other is hidden in the above form of inference but is equally unreliable. If this is right, then inferences of the above form should sometimes lead to false conclusions *even when the LA is valid*, because of the failure of this other assumption.

Here is an example. One day a caller to National Public Radio's Car Talk radio call-in program informed the experts, Click and Clack the Tappet brothers, that she had a problem with her accelerator. When asked to elaborate, she said that the car worked normally in all respects except that it had difficulty accelerating. Click and Clack were of course rather amused and informed her that cars do not have accelerators (as opposed to accelerator pedals); the problem was, most likely, with her carburetor.

Now, the LA is, by and large, true of car engines. Not only do most components interact directly with only one or a few other components; if there is damage to a component, undamaged components continue to function normally. Yet the inference from selective deficit (poor accelerating) to dedicated component (accelerator) still fails. The caller was relying on a further hidden assumption: that each identifiable function of the car must be accounted for by a dedicated component. It was this assumption, not the LA, which led to her mistaken conclusion that she had an accelerator which was malfunctioning. This assumption can be formulated the following way:

Premise 3 (The other locality assumption). For a given specifiable ability of a system there must be a component of the functional architecture responsible for that ability.

Without this assumption, the inference from a selective deficit in ability A to a component dedicated to A simply fails to go through, since the ability A may be one that corresponds only to the operation of many components (which may each have only local effects). If this assumption is false or unreliable, it will lead to unreliable conclusions.

When neuropsychologists make inferences of the kind Farah rightly criticizes, they are typically making both locality assumptions. Farah's models and arguments successfully demonstrate that both assumptions are unreliable – and probably false in the case of the brain. The twin locality assumptions are at the very heart of the localizationist tradition. In undermining them both, Farah is truly playing Flourens to Fodor's Gall.

Prosopagnosia, conscious awareness and the interactive brain

Robert Van Gulick

Department of Philosophy, Syracuse University, Syracuse, NY 13244-1170
Electronic mail: rrvangul@sunrise.bitnet

Farah makes a convincing case against the methodological necessity, if not against the truth, of the locality assumption. In each of the three cases she considers, she provides a serious and plausible alternative to the standard explanations without relying on assumptions of locality or Fodor-style modularity. Since, as she rightly observes, both sets of inferences fit the existing data, the choice between the competing models must be based either on further data or on other empirical considerations, such as simplicity, or coherence with established principles of brain organization. In the latter regard, she finds her alternative nonlocalist explanations superior to their standard competitors. However, the sorts of considerations she invokes (sect. 3.1) weigh rather lightly on the empirical scales. With respect to both visual attention and prosopagnosia she appeals to the fact

that her model has fewer components. Though parsimony counts for something, it is a far from reliable guide to truth. Reality unfortunately often turns out to be more complicated than we would prefer, and nature's designs in brain organization, as elsewhere, are often far from ideally simple. Thus the truth, as opposed to the necessity of the locality assumption, remains very much an open question.

Several special questions arise with respect to Farah's analysis of prosopagnosia. She notes that at least some researchers (de Haan et al. 1992) relying on the locality assumption have attempted to explain the dissociation between covert and overt face-recognition tasks as the result of damage to the conscious awareness system that prevents it from accessing the results of the face-processing module, which itself continues to proceed normally in covert recognition. Farah claims that in her alternative model the face components, when not damaged, do the work of both the face-processing module and the conscious awareness system in the standard model, and that when they are damaged they, by themselves, replicate the clinical pattern of severely diminished performance on overt recognition tasks combined with partially preserved covert performance. Most important, Farah states that the ability of the model to account for this dissociation "depends critically on this violation of the locality assumption" (sect. 2.3.4).

However, there are at least four problems with Farah's position. First, the inference she is criticizing is open to objection quite independently of any challenge to the locality assumption. Inferences based on single dissociations are notoriously risky. The preservation of ability *A* in conjunction with the loss of a different ability *B* can at best be taken as weak evidence that abilities *A* and *B* depend on distinct components, even if one is assuming localism. There is always the rival hypothesis that *A* and *B* depend on the same components but *B* requires a higher level of functioning of one or more of those components than *A* does. In its damaged state the relevant shared component is simply no longer able to function at the level required for the more demanding *B*-type tasks (Shallice 1988). Double dissociations in which the converse patterns of loss and preservation are found in different patients thus provide much better evidence that the abilities depend at least in part on distinct components.

Second, it is this very inferential weakness that is exploited in Farah's criticism of de Haan et al. In her alternative model, the face units, which are close to constituting a component in the more standard sense, are damaged to such a degree that they are no longer able to meet the demands of overt recognition but still function well enough to meet the lesser requirements of covert recognition. Farah's model seems to be on a par with the sort of resource limitation hypothesis that would (or at least should) be standardly entertained as an alternative explanation of single dissociations by those working within the locality assumption.

Third, Farah's final claim is thus called into question. In what sense does her explanation depend critically on violation of the locality assumption? It would seem it does not. Admittedly, the preservation of some residual function in the faces unit is explained by the distributed nature of the PDP-type representation it uses, but that is a different issue. The locality issue in dispute concerns not whether information is represented in a local or distributed fashion within components, but whether the other components of the system themselves continue to function in the same way after damage to a single component as they did before. I fail to see how Farah's model of prosopagnosia requires violation of the locality assumption in the latter sense. Consider relearning after damage. Clearly, what is relevant is that even in its damaged state the faces component retains enough of its earlier distributed representation to allow for its reconstruction with training more quickly than it could if it had retained no portion of that representation. However, this all concerns what happens *within the faces component itself*; it does

not involve interactively induced changes in the functioning of other components.

Fourth and last, it is not clear that Farah's model in its intact state accounts for all the abilities of normal subjects. Remember that normal subjects report introspective and conscious awareness of recognizing familiar faces. Those reports and the conscious awareness that underlies them need to be explained, as does their joint absence in prosopagnosics. A network's ability to match a face with a name or an occupation is not the same as having a conscious experience of recognizing the person. Farah's model is of course not designed to simulate conscious awareness, nor does it do so. But a fully explanatory account of face recognition does at some point have to address that phenomenon. Normal subjects do consciously experience familiar faces as familiar, and that fact needs to be explained.

In this context it may be helpful to think about conscious awareness in a way that is in keeping with the interactive spirit of Farah's approach. In the model she criticizes, the conscious awareness system is treated at least implicitly as yet another module, a distinct component, to which inputs are passed by the face-processing module. However, there are good reasons to doubt the existence of any privileged location or unit of the brain serving as the conscious awareness module or what Dennett and Kinsbourne have called the "Cartesian Theater" (1992). The formation of conscious states is more likely to involve the integration of local representations from many specialized brain regions into an interactively unified representational state, what Kinsbourne refers to as "an integrated field theory of consciousness" (1988; 1993). Local representations are not passed on to any supermodule; they are "recruited" for incorporation into global coherent unified states. One of the functions of such states would be to bring isolated items of information, such as the visually processed image of a face, into contact with a wide range of other items of information in a way that subserves the sort of flexible and relatively open-ended range of behavioral responses we typically associate with conscious awareness. Such a theory might well be combined with Farah's model, in which the role of a given representational subsystem can shift depending on which other systems it is interacting with. Moreover, the lack of conscious face recognition in prosopagnosics might result from the fact that the residual representations after damage are now too attenuated to be recruited into preexisting global patterns, although with retraining the relevant threshold might be achieved again soon – or at least more quickly than during initial learning. Such interactive and global theories of conscious awareness are admittedly speculative, but conscious awareness does eventually require explanation, and such theories seem to show a natural affinity with the interactive models Farah proposes.

The symbolic brain or the invisible hand?

René van Hezewijk and Edward H. F. de Haan

Department of Psychonomics, Utrecht University, 3584 CS Utrecht, The Netherlands

Electronic mail: hezewijk@fsw.ruu.nl; dehaan@fsw.ruu.nl

Not even a connectionist can do without some sort of computational (or competence) theory (Marr 1982) of what it is that the brain must do, wherever or however it does it. In this commentary, we discuss first whether parallel distributed processing (PDP) as a general conceptual framework can do a better job than the old-fashioned symbolic models. We will then address two problems with the specific PDP implementations of the living-nonliving dichotomy and covert face recognition.

The locality assumption in neuropsychology borrows partly from common sense and partly from sophisticated symbolist

theories of what a person's competences should be. When a particular competence has become disrupted after neurological disease, the search for the functional (and physiological or anatomical) nature of the dysfunction begins. Connectionist explanations for dysfunction are either parasitic on the common-sense and symbolist characterizations of cognitive (dys)function or must find a way to derive from their brain-style models what the impairment could be. The latter (independent) way is the hard – if not impossible – way. We have never seen an example. So the first option remains. In our view, this means that the PDP approach, interesting as it is, must drop its pretense of being a completely independent alternative conceptual framework. It cannot be much more than an "implementational theory" of cognitive (dys)function, complementary to competence theories. Some of the consequential claims will have to be toned down as well; for example, it remains an empirical question whether the connectionist idealizations are supported by physiological and anatomical research. Also, if assumptions such as the locality assumption are falsified, this has repercussions for PDP modeling when it is embedded in the symbolic approach. Finally, conclusions concerning PDP models are "local" too, in the sense of not predicting more dysfunctions or different ones from those the symbolic models predict. In other words, unless connectionist theory (PDP-style) independently infers cognitive (dys)functions from changes at the brain level, it will produce only *ad hoc* explanations of symbolically defined cognitive dysfunctions (for the precise meanings of *ad hoc* see Lakatos 1970).

Our second point is that PDP models do not seem to be able to distinguish between *impaired* networks (lesioned models) on the one hand, and networks that have learned *impaired behavior* on the other. One might just as well program a PDP model to produce the same performance as that observed by Farah and McClelland (1991), Farah et al. (1991), and Farah et al. (1993), for this would clearly not result from damage but from learning to produce the desired results. This is the consequence of the "programmatically ad hocness" one often finds in PDP-style modeling. PDP lacks clear theoretical answers (logically prior to experimenting with and calibrating the network) to the following questions: (1) How many units does one need? (2) How many layers does one use? (3) What learning rule does one use? (4) How many trials does it take to have the network learn the job it has to do? (5) Who or what defines the task? And in the case of lesioning the model: (6) Which units are damaged, and (7) how many units? The answer to the fifth question is found outside the PDP program itself and lies in the symbolic approach. The other questions were not answered by Farah in the present article and cannot be answered by the reader by inferring from the (PDP) theory. We suggest, and will believe until shown otherwise, that this is left to the discretion of the network programmer, the "invisible hand" in PDP-style modeling.

This is not to say that there are no *ad hoc* practices in the "box-and-arrow approach." However, in this approach it is customary, if not obligatory, to test an explanation with new predictions for both impaired and unimpaired subjects. All we can see in the reports of PDP models is how anything done by the traditional approach can be done with the PDP approach as well. Regarding the empiricist claim for "sufficiency of theory," Farah is of course right that no theory can ever do more than fit the known data at the time. However, what happens next? It seems to us that PDP modelers just ask themselves: "What is the next phenomenon to model?" whereas (realist) symbolic modelers would continue: "So what are the predictions from this model and how can I test them?" In this sense, the PDP approach can be characterized as "abortive," whereas the symbolic approach has inherent "excess content" (Popper 1963; Lakatos 1970). The locality assumption may be a weak spot of the classical symbolist approach; however, the "keep it local and address one issue (defined symbolically) at a time" assumption is a weak spot in PDP-style connectionism.

Our third point is that the lesioned PDP models used by Farah to demonstrate nonlocality should be tested for recovery, be it spontaneous, supervised, or by deviation. Some patients suffering from the dysfunctions Farah discusses recover, while others do not. We wonder whether the lesioned models show certain patterns of recovery. A related question concerns whether, assuming that PDP models are able to describe the observed neuropsychological data without using the locality assumption, they do a better job overall. For example, one of the great triumphs of PDP-style modeling is the "graceful degradation" that damaged models show; but is there always (only) graceful degradation in the neuropsychological reality?

Our fourth point concerns the assumptions necessary to make the neuronal idealizations of the PDP approach possible. Inhibitory and activating connections are realized in the same connections in PDP models. But antidromic connections do not exist in the real neural anatomy. Also, the often used learning strategy of backpropagation is biologically implausible. Grossberg (1987) has shown how complicated the necessary neural hardware is that is needed to implement backpropagation (but see Davis et al. 1989, for an alternative viewpoint). PDP-style modeling uses a sweeping simplification of nature. Therefore the PDP model shown in Farah's Figure 7 is not as simple as suggested.

A fifth and last general point is the following. Farah states that "the effects of damaging one component should be relatively local." This assumes the rather rough and unsophisticated version of modularity of Fodor (1983). A more sophisticated interpretation leads to different predictions. For example, Jackendoff (1987) suggests in his intermediate level theory (ILT) that informational encapsulation is relative to a certain representational level that is more refined and better defined than Fodor's, and that informational encapsulation need not correspond to strict cerebral encapsulation (= localization) of processes. For example, one would predict that if there are relatively local impairments somewhere in "lower vision" (e.g., in those neuronal substrates that compute primal sketches), this only affects higher vision (e.g., 2½D sketch, the spatial or 3D model, conceptual structure) insofar as the *information* used at the higher level originates in the lower level. The symbolist approach, for example, Jackendoff's ILT, makes it possible to predict the kinds of interactions that can be found between different representational levels in the more sophisticated sense. So the real issue is not only the *degree* of interactivity or nonlocality but also the *nature* (or content) of the interactions. PDP-style modeling and theorizing does not have much to offer in this sense.

Finally, we would like to focus on two of the three specific implementations used by Farah to argue her position in the target article. First, in her discussion of the selective impairments in knowledge about living and nonliving things she states that the PDP model can accommodate the existing neuropsychological data. We are curious to know whether this also includes the findings of Young et al. (1989), who found that the patient MS was severely impaired in overtly retrieving knowledge about living things compared to nonliving objects, but that this difference disappeared when his knowledge was probed in an indirect priming experiment. Are we right that this might require the "invisible hand" of the programmer, or can the model cope as it stands?

De Haan et al. (1992) were using the concept of "modules" (quotation marks in the original) to describe the dissociation between autonomic responding, information processing, and conscious awareness. Their patient demonstrated covert face-recognition effects on both physiological and behavioral indices in the absence of consciously acknowledged recognition of faces. The rather unsophisticated use of modular organization was intended to describe this puzzling phenomenon. We are very keen to learn how the dissociation from *awareness* can be modeled in a distributed network. Is it the ghost in the PDP machine?

The localization/distribution distinction in neuropsychology is related to the isomorphism/multiple meaning distinction in cell electrophysiology

Gerald S. Wasserman

Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907-1364

Electronic mail: codelab@psych.purdue.edu

I wish to offer here the "bottom up" observation that the neuropsychological concepts that form the basis of Farah's excellent target article have their counterparts in concepts now percolating up through neuropsychology's cellular substrate. There are two links: on the one hand, if one holds that a particular function is localized to a particular place in the brain, then one must also hold that the cells resident in that place can only be involved in supporting that function. On the other hand, if one holds that functions are distributed to many places around the brain, one must also hold that any given place in the brain will tend to mediate more than one function. Hence, the distribution assumption would lead one to expect single cells to support more than one function.

The latter half of this proposition was clearly articulated by Farah, using parallel distributed processing (PDP) terminology, in section 1.4 of her target article, where she wrote: "Different entities can therefore be represented using the same set of units, because the pattern of activation over the units will be distinctive."

The fact that certain properties are exhibited by PDP models, however, does not necessarily limit what goes on in the real nervous system. The present commentary therefore draws attention to the fact that investigations of the cellular elements which collectively constitute the various "places" of neuropsychology have recently led to serious discussions of this issue on a microscopic level: Specifically, do real individual nerve cells contain information which relates to one function or to several? Elsewhere (Wasserman 1992), I have provided an extensive review¹ of this cellular discussion. What follows is a brief summary of that review, omitting citations.

Isomorphism. The cellular counterpart of the locality assumption has been named the isomorphism hypothesis. It is generally introduced in standard texts in the form of a laconic ostensive definition focusing on the special case of spikes propagating in an axon. Discussion is often narrowly couched in terms of "frequency coding," with the frequency of axonal spikes the index of information resident in that axon. This amounts to a tacit assertion that an axon conveys information only about one thing.

This is an old view which was first clearly stated when the advent of electronic amplification made it possible to record from single nerve cells. That technology forced a consideration of the antecedent question posed by the fact that cellular sensory responses differ in form from the stimuli which evoked them. In conjunction with certain psychobiological correlations, this led to the conclusion that sensations were isomorphic with the waveforms of cellular sensory responses. The original meaning was that the magnitude of a sensation was indexed by the magnitude of the cellular responses and the time course of a sensation was indexed by the time course of the cellular response.

The isomorphism concept developed over the years by incorporating more recent discoveries. Of particular importance were the manifold indications that neurons often do not respond unless an appropriate spatiotemporal pattern of stimulation is delivered. The isomorphism concept thereby became generalized to the concept that high activity in a given nerve cell was a strong indication that one particular pattern of stimulation was present.

The various manifestations of the isomorphism hypothesis share the notion that a nerve cell is a univariate device whose

activity signals some one thing of particular functional significance. The more the cell responds, the more of that function is present. It then becomes the business of the investigator to determine the role of a given cell by inquiring into the nature of the function signaled by its level of activity.

Patterning in neuronal signals. Problems developed fairly early in the history of cellular electrophysiology, however, when investigators tried to fit the complexity of real neural responses into the Procrustean bed of the isomorphism notion. One of the earliest findings, for example, is that changing wavelength changes the pattern of response in some color-coding neurons. Even today, appreciation of the deeper significance of this well-known fact is often blunted by characterizations of these complex patterns which constrain them to fit into a univariate metric.

Over time, many complex response patterns have been characterized in many neurons, and these patterns have been demonstrated to be under the influence of multiple sensory variables. This suggested that the pattern of a nerve cell's response carries information that is not encoded in its overall activity level.

Similar problems became evident on the output side. It became abundantly clear – from studies of synaptic transfer in both muscles and neurons – that the efficacy of transmission depended critically on the pattern of presynaptic activity as well as on its quantity.

Multiple meanings. These problems gradually led many scholars to consider the possibility that the complex pattern of a neuron's activity might be capable of signaling information about more than one aspect of sensory input or motor output. It is now not uncommon for this notion to be expressed in terms of the "multiple meanings" that may reside in a single cell's activity.

At first, this discussion was primarily formal and hypothetical, with a limited empirical base. More recently, careful quantitative investigations have demonstrated that multivariate properties are very common in neurons. Particularly interesting are reexaminations of certain classical findings of visual physiology, which have been illuminated by multiple-meaning theory. Both the contrast-sensitive receptive fields of retinal cells and the orientation-sensitive receptive fields of cortical neurons have been thoroughly scrutinized in this way. At both levels of the visual system it has been demonstrated that important information about the stimulus resides in the temporal waveform of a cell's activity. Because of this, simple univariate assessments of the overall responsiveness of such cells will necessarily produce ambiguous results.

Pattern predicate. This interpretation of the data predicates that the nervous system must be able both to create and to process structured patterns in neuronal responses. This suggestion has led to very interesting results. Two may be of special significance to neuropsychology. First, neural mechanisms appear to exist which can convert the temporal pattern of activity occurring in a single part of one neuron into a spatial pattern of activity with differential effects appearing in different places. Because of this finding, bare evidence of connectivity may be profoundly incomplete. Second, consciousness itself – as indexed by the effects of anesthetics on neuronal responses – may depend on the pattern of activity in neurons, not on their activity level.

These recent cellular findings will surely influence the conceptual nervous system on which much of neuropsychology has been erected. Even if the strongest postulates of multiple meaning theory, as reviewed in Wasserman (1992), are ultimately subject to extensive revision, neuropsychologists interested in the locality issue may want to consider the evidence that has recently been uncovered by these seminal cellular investigations.

NOTE

1. Wasserman (1992) is also available by e-mail from the connectionists archive by anonymous file transfer protocol (ftp). To retrieve it, log on and do the following (where % stands for your own system's prompt):

```
%ftp archive.cis.ohio-state.edu
Name: anonymous
Password: your e-mail address
ftp> cd pub/neuroprose
ftp> binary
ftp> get wasserman.mult_mean.ps.Z
ftp> quit
%uncompress wasserman.mult_mean.ps.Z
%!pr -s wasserman.mult_mean.ps
```

If the printer for this job resides on a remote machine, this large (i.e., graphics-intensive) file may require that an operator issue the print command directly from the remote console.

What counts as local?

Andrew W. Young

Department of Psychology, Science Laboratories, University of Durham, Durham DH1 3LE, England
 Electronic mail: andy.young@mrc-apu.cam.ac.uk

Selective impairments have attracted great interest because they provide a powerful source of evidence for testing psychological theories. The argument is that an adequate psychological theory of how a particular ability is organised should be able to account for the patterns of impairment found after brain injury. To the extent that it fails to do this, the theory should be revised or abandoned. The force of this approach is clearly seen in Warrington and Shallice's (1969) finding of preserved long-term memory despite poor immediate recall, which is incompatible with the view that perceived stimuli must pass through a period of short-term storage before they can enter long-term memory, and in Marshall and Newcombe's (1966) demonstration of semantic reading errors ("ill" read as "sick," etc.), which imply that reading cannot be exclusively mediated through recoding print into sound.

Since these pioneering investigations, neuropsychological studies have often been used both to test and to derive models of the functional architecture of human cognition using an approach widely described in terms of a metaphor of brain injury as a cruel and somewhat haphazard natural experiment that can occasionally carve nature at its mental joints.

Farah argues that interpreting deficits in this way involves a "locality assumption," in which "the removal of one component would have only very local effects on the functioning of the system as a whole." She invites us to think this "naive," suggesting that it has led to "a mindless reification of deficits."

Those of a historical bent will recognise that we have been here before; the work of the nineteenth-century diagram makers was dismissed by people who did not accept that particular patterns of deficit could be attributed to the loss of particular cortical centres, or even that there were distinct patterns of deficit at all (McCarthy & Warrington 1990; Shallice 1988). Eventually, this globalist critique was recognised as overstated, and the localist approach was revived with more stringent standards of inference and evidence.

A crucial problem is that "local" is never defined in the target article. Yet what counts as local depends entirely on one's point of view. To a person on Mars, or even in London, Durham might well be in the locality of Sunderland, but it does not seem so from here.

This is important because, to the best of our knowledge, the locality assumption is indeed appropriate at what might be considered coarse deficit scales. For example, I know of no evidence to suggest that visual deficits are inevitably accompanied by hearing loss, or that receptive aphasias necessarily create problems in moving your toes. The fact that neurons are highly interconnected influences this no more than the interaction of the molecules in the atmosphere implies that when I

breathe here in Durham any consequent breeze will be detectable in Sunderland. Farah would seem to agree, since she comments that "different brain areas are dedicated to representing information from specific sensory and motor channels." This is locality without invoking the label.

Since locality already applies to coarsely defined deficits, the issue to be resolved only concerns the level at which it might no longer be useful because evidence of interactions between fine-grained deficits starts to emerge. Yet the target article treats this issue as if it could be determined by theoretical fiat, and as if the adoption of parallel distributed processing (PDP) models somehow settled the necessary theoretical choices. It does not.

Consider Figure 1. This is the first PDP model of an entire personality; SAMSON. Conveniently, we happen to know quite a lot about its behaviour because it is the same model as Figure 11 in the target article, with the labels changed. In this model, we can simulate the effects of shaving Samson's head by reducing the activations of the hair length units. Notice that when we do this the model's strength begins to fall and it can no longer engage in Delilah-lusting or temple-shattering behaviours. Yet if the hair length units can subsequently gain in activation, its strength returns.

The point is this: connectionist models provide a powerful tool for implementing and exploring potential accounts of certain neuropsychological findings, and especially certain types of dissociation. They do not substitute for evidence, however. It is evidence which must determine the correct choice of components to be modelled, and evidence which will ultimately determine the extent to which one can rely on locally based descriptions of mental abilities. A simulation can only be judged by the extent to which it is compatible with a broad range of existing evidence and capable of generating predictions to be tested.

As SAMSON shows, the level of description at which an interpretation is attempted is as crucial in simulation as it is in neuropsychology. There is a risk of overinterpreting what computer models are doing and losing sight of difficult questions, as Searle (1984; 1992) has pointed out. We so easily slip into reading our own ideas about attending into the boxes labelled "attention" in Figure 7 of the target article. Similarly, the model shown in Figure 11 does not really recognise faces, and it isn't conscious either. What it does is to show that impairing a particular type of system can leave it able to perform some functions and not others. This is certainly a pertinent observation, but the same point has already been made by researchers who have seen it as a modest step forward, not a solution (Burton et al. 1991). For example, a striking feature of prosopagnosic patients who show covert recognition of familiar faces is that they do not seem to act on this in everyday life; they

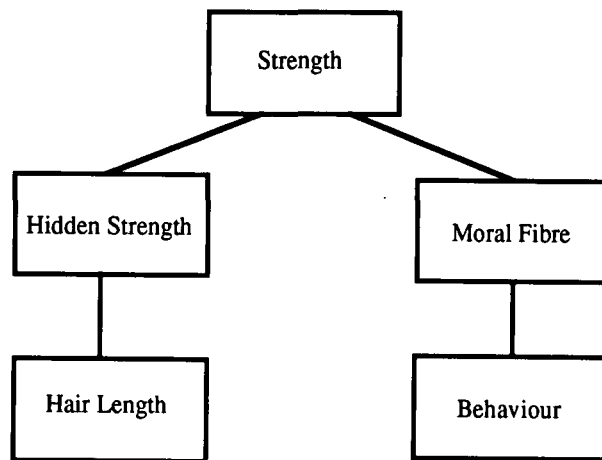


Figure 1 (Young). SAMSON (Simulation of Alopecia and Masculine Strength with an Organised Network).

do not greet familiar people in the street without knowing why. Such observations are consistent with the commonsense conception that conscious recognition is involved in intentional actions, and one wants an account of how this happens, yet the type of model shown in Figure 11 has little to say about this.

Locality, then, is an empirical issue, not an overriding assumption, and what is to count as local depends on some notion of the appropriate scale of theory required. Connectionist models can have many elements, but at present the brain has far more. A degree of interactivity of deficits is to be expected, but this may only happen at finer-grained levels of analysis than the target article implies.

Modularity need not imply locality: Damaged modules can have nonlocal effects

Edgar Zurif^a and David Swinney^b

^aDepartment of Psychology, Brandeis University, Waltham, MA 02254 and Aphasia Research Center, Boston University School of Medicine;

^bDepartment of Psychology, University of California, San Diego, La Jolla, CA 92093

Electronic mail: zurif@brandeis.bitnet

Farah takes a skeptical look at several accounts in which behavioral deficits are taken to reveal the local effects of disrupted cognitive components. We agree with her on the inadequacy of these accounts and think that in each case her interactionist descriptions may well be better. However, we disagree with her view that the accounts she criticizes are based on a modular cognitive architecture of the sort proposed by Fodor (1983; and she seems unaware of the existing empirical tests for modularity). We especially disagree with her argument that encapsulated modules are not useful in neuropsychological explanation because they do not interact enough.

Contrary to Farah's views on the matter, modular components of Fodor's sort are not necessarily less interactive than those that figure in maximally interactive PDP frameworks. The difference between interactive and modular systems turns, not on the amount of interaction, but on *when* interactions among constituent processes take place. In the interaction account, interactions seem to occur as available, whenever potentially useful; in the modular account, they occur only at the endpoints of the constituent computations (Fodor 1983; Fodor et al. 1992; Forster 1979; Garrett 1981). This being so, the effect of a damaged process can be as nonlocal or as local in a modular system as in an interactive one – it depends on the overall layout of the components, how they share resources, and so on (see, e.g., Forster 1979).

Farah has it otherwise, however. She stipulates that if interactions occur only at endpoints, they must be limited. To justify her claim, she aims her nonlocality arguments at targets that, as we have already suggested, are rather flimsy (or at least they seem that way in Farah's accounts of them).

Consider the face-recognition case. The finding is that patients can recognize faces without being aware that they have done so. As Farah states it, the locality position here is that a functional lesion outside the face-processing module prevents the module from transferring its product to conscious awareness. This is hardly a test of locality or modularity, however; all it does is restate the phenomenon.

Farah, by contrast, puts some details into her interactive model of face recognition, but the contrast she provides is not between an interactive and a modular account. Rather, it's between an interactive account and one that fails altogether to consider the details (and the time course) of possible interactions – fails even to provide the necessary ingredients for thinking about interactions of the sort considered by Farah.

As for the semantic memory case, the partitions entered for the locality position are very incompletely drawn. The living-

nonliving distinction has little organizational impact; generic concepts that serve as unique beginners of separate hierarchies seem, at the least, to distinguish plants, animals, and persons (Miller & Fellbaum 1991). And the effort to transform the living-nonliving distinction into an equally unalterable one between sensory and functional information does not work. As Farah points out, the two sets of categories are simply not coextensive. And neither set seems to bear the burden of current theoretical analyses of lexical conceptual structure – analyses that seek to tie together conceptual structure and grammar and that test semantic distinctions against large amounts of lexical data, in short, analyses that go beyond simple intuitions about defining a word (Levin & Pinker 1991).

More pointedly, however, the partitions in the so-called locality model of lexical semantics seem to define modules only in the weak sense that Farah claims to have no quarrel with; the modules are nothing other than expressions of the fact that different kinds of knowledge are accommodated in the lexicon. These partitions do not seem to require the property of cognitive impenetrability. It's one thing to create a partition to account for relatively selective naming impairments, quite another to show that each of the categories the partition defines is encapsulated and has only minimal interactions with the rest of the system. These are empirical matters that do not seem to have been settled by Warrington and her colleagues.

There are, however, instances in which disruptions to encapsulated modules have been isolated, and these disruptions have been shown to have distinctly *nonlocal* consequences. What we have in mind turns on characteristics of lexical activation during the course of sentence comprehension: immediately upon hearing a polysemous word in a sentence, the normal (neurologically intact) listener activates all of that word's meanings, not just the one relevant to the sentence context. Only after a time delay of approximately a second does context exert its effect; only after that time does context damp all of the word's meanings save the relevant one (e.g., Swinney 1979). The general lesson here is that sentence processing does not consist of a set of maximally interactive processes. Rather, the lexical activation device seems momentarily informationally encapsulated – in short, it behaves as a module.

More directly to the present point, lexical activation follows this modular course in Wernicke's aphasia too, but not in Broca's aphasia (Swinney et al. 1989). It is not that the lesion that underlies Broca's aphasia causes "unencapsulation." Rather, it seems to cause a "module-internal" problem – the Broca's patients remain uninfluenced by sentence context but appear to activate word meanings in a slower-than-normal fashion (Swinney et al. 1989; also Prather et al. 1991).

In an effort to determine how this modular disruption might ramify within the comprehension system, we have lately focused on the real-time syntactic formation of within-sentence dependency relations – on the linking of antecedents and anaphors (including here the reactivation of moved lexical elements at the site of their extraction). This syntactically directed lexical reactivation must be implemented under strict time constraints (immediately upon encountering the structural licensing conditions; Swinney & Fodor 1989; Swinney & Osterhout 1990). The results under these circumstances are as expected: the Broca's patients fail to reactivate lexical elements within the normal time frame – they are unable, that is, to establish syntactically indicated dependency relations in a normal manner (Zurif et al., in press). Moreover, this lexical reactivation problem shown "on-line" connects directly to independent, "off-line" analyses of *sentence* comprehension limitations in Broca's aphasia: it accounts for the repeated observation that comprehension is particularly difficult for these patients when one element in a sentence must be interpreted with respect to another element in that sentence (e.g., Ansell & Flowers 1982; Caplan & Futter 1986; Caramazza & Zurif 1976; Grodzinsky 1986). The more general point, of course, is that

within a modular system, disruption to one module can have far-reaching (nonlocal) consequences.

The above particulars hold only for Broca's aphasic patients. In line with their normal lexical activation functions, Wernicke's patients show normal patterns of syntactically governed lexical reactivation. That is, they establish intrasentence dependency relations within the normal time frame (Zurif et al., in press). So to the extent that Broca's and Wernicke's aphasia implicate different lesion sites, the modular lexical activation disruption and its specific syntactic consequences have lesion localizing value. And this is another feature of a modular system (Fodor 1983). What also warrants emphasis is that localizing signs of this sort emerge most clearly in "on-line" analyses (see Zurif & Swinney, in press, for a detailed discussion of this last point).

In any event, Farah's test-case "modules" seem to be isolated as a means of redescribing patterns of sparing and loss following brain damage, and that does not guarantee the linchpin characteristic of a module, namely, its encapsulation. Evidence for encapsulation should emerge as a result of charting mandatory and fixed operating characteristics of one or another part of a system. From a processing perspective – and since she cites Fodor on the definition of a module, this seems to be Farah's perspective as well – a module is informationally encapsulated only over the time course of its operation. Whether modules of this sort have local or widespread effects when damaged is a matter that should not be stipulated in advance of actually measuring the consequences of their disruption.

Author's Response

Interactions on the interactive brain

Martha J. Farah

Department of Psychology, University of Pennsylvania, Philadelphia, PA
19104-6196

Electronic mail: mfarah@cattell.psych.upenn.edu

R1. Modularity, the locality assumption and neuropsychology

R1.1. Implications for modularity. Many of the commentaries touched on the issue of modularity. Four main points were made. First, several commentators noted that my alternatives to modularity are themselves modular, in the sense of having components that represent different kinds of information. For example, **Servos & Olds** state that I seem both to reject and accept the concept of modularity; they quote my statement that "it is well known that different brain areas are dedicated to representing information from specific sensory and motor channels." **Young** quotes this same passage and says "This is locality without invoking the label." **Butterworth** also notes the apparent inconsistency and asks whether I believe there are "regional specialties" or just undifferentiated "mass action." The confusion stems from two different meanings of the word "modular," which I distinguished in section 3.3.1. The first meaning is "informationally encapsulated," and it is this sense of modularity that is challenged by failures of the locality assumption. The second meaning is "specialized processors," for example, those subserving visual and functional knowledge,

and this is not challenged by failures of the locality assumption. Of course, as **Glymour** points out in the second section of his commentary, to the extent that different modules influence each other, the nature of the information they represent becomes less distinct. Less distinct does not mean indistinct, however, and distinctions that are a matter of degree are not necessarily vague.

Second, it was pointed out by several commentators that Fodor (1983) restricted his claims of modularity to peripheral systems, among which he included, for example, vision and language (**Butterworth, Chater, Kinsbourne, Umiltà, Zurif & Swinney**), whereas my discussion of modularity is not so restricted. This is true. Fodor's main defining *criterion* for modularity, informational encapsulation, describes the architectures for which the locality assumption holds; it was not my intention to discuss his substantive claims about which systems will and will not conform to that criterion. Although I think this is reasonably clear in the introduction (sect. 1.1), I admit that elsewhere I slipped from what should have been "the hypothesis that Fodor's definition of modularity holds for the cognitive architecture" to "Fodor's modularity hypothesis." **Zurif & Swinney** offer a further criticism of my account of Fodorian modularity, claiming that the difference between modular and interactive systems turns, not on the amount of interaction, but on *when* interactions among constituent processes take place. This is not consistent with my reading of Fodor (1983), for example, his statement that in informationally encapsulated systems, "of all the information that might, in principle, bear upon a problem . . . only a portion (perhaps only quite a small and stereotyped portion) is actually admitted for consideration" (p. 70). Perhaps our views can be reconciled by noting that holding back sources of information that could in principle be used by a processing component until that component has completed its computations is not different from simply holding them back forever, as far as that component's behavior (input-output function) is concerned.

A third point concerned the scope and limits of informational encapsulation. **Glymour** formulates the issue clearly, distinguishing among a number of alternative claims that could be made concerning the encapsulation of modules and consequent locality of the effects of damage. There could be locality for the effects of damage to any one component on any other component (his Equations 1 and 2), or for the effects of damage to particular components on particular other components (Equations 3 and 4), or for the effects of damage to any component on the functioning of a particular component (Equations 5 and 6). Furthermore, the locality or non-locality of the effects of damage could be measured relative to the direct input to a nondamaged component (odd-numbered equations) or relative to the input to the system as a whole from the environment (even-numbered equations).

Which of these is the version of locality that I discuss? Although **Glymour** suggests it might be (1), it is none of the odd-numbered possibilities. A violation of this type of locality would involve a component changing its input-output function, and as **Oaksford** and **Plaut** both point out, this does not occur in any of the examples I discuss. Nor, I might add, would one ever expect such a change on the grounds of parallel distributed processing (PDF) or

any other existing proposals for the nature of human information processing. Plaut phrases this distinction between odd- and even-numbered possibilities in terms of the "behavior" (output to other components) versus the "computation" (output given particular inputs) of a component, and correctly states that "clearly the nondamaged portions of the networks compute normally but behave abnormally in response to corrupted input from damaged portions."

Glymour claims my arguments apply only to his version (2), which would indeed be unfortunate if true, because it requires absolutely no communication among components, a nonsensical extreme of informational encapsulation. **Grodzinsky & Hadar** and **Zurif & Swinney** also point out that even strong modularity requires some intermodule communication if the system is to do anything. I agree, and the distinction proposed in the target article is between architectures in which there is abundant versus minimal intercomponent communication (e.g., sects. 1.1, 1.3). This quantitative distinction is important for the kinds of qualitative inferences we draw about the cognitive architecture from patients' behavior. I will review the reason for this claim in the context of **Glymour's** discussion of the **Farah and McClelland (1991)** model of semantic memory.

Glymour correctly points out that the violated version of the locality assumption in this case is of type (4), but he says that this is not interesting for two reasons. Let me first dispense with the second reason, that the behavior of the model is "obvious from [its] functional structure." The same point was raised as a criticism by **Diederich**, who complained that "the effect of damage to the networks can easily be explained by the network structures." This line of criticism strikes me as bizarre, because it suggests that it is not desirable for theories to explain phenomena too plainly. It is also ironic in the present case because PDP models have recently been criticized on the grounds that, although they may account for the data, the way they do so is frequently obscure (**McCloskey 1991**).

As his first reason for finding violations of type (4) locality uninteresting, **Glymour** states that no one held contrary views about the encapsulation of the particular modules in question. If this were true, then I agree that this example would not represent a failure of locality in any interesting sense. However, the most straightforward account of how functional semantic information is accessed through a verbal query involves communication between verbal systems and functional semantics, not communication between visual and functional semantics. This is the belief concerning encapsulation that is violated in our account. In this sense the account is contrary to accepted or default views.

A fourth point in the commentaries concerning modularity is that modular architectures may also exhibit nonlocal effects of local damage. There are two ways that this might happen. One way is discussed by **Plaut** and is closely related to the previous point, that informational encapsulation is a matter of degree. **Plaut** explains that components will behave abnormally, given abnormal input, whether they are in a modular or an interactive architecture. Of course, components will less often receive abnormal input following local damage in modular architectures, because in general their components have relatively few sources of input. **Plaut** questions whether

the locality versus nonlocality of the effects of brain damage per se is the key distinction that neuropsychologists should consider, as opposed to interactive versus modular architectures. These appear to me to be different formulations of the same distinction, and in fact the title and text of the target article address both. Each formulation has the disadvantage that it requires us to rule out a silly version: the "silly locality assumption" is violated by any change in behavior of any nondamaged component, including the relatively small number of downstream components in a modular architecture. The corresponding "silly modularity hypothesis" is that components are not just encapsulated from most other components, but from all other components.

In contrast to effects of local damage on the small number of downstream components in a modular architecture, **Semenza** and **Chater** point out that such an architecture could manifest nonlocal effects of damage by compensatory strategy shifts. **Chater** offers the example of a subject who, after losing his lexical route to reading, might rely on phonological and semantic routes that were not previously used. In such a case the output of nondamaged components is not changed; rather, the components whose outputs control behavior have changed. To the extent that we have prior knowledge of the components involved (as we do with semantics and phonology) and to the extent that our methods allow us to identify the components whose outputs are controlling behavior (which, again, is the case for this example), this type of nonlocal effect will be more tractable within an assumed modular framework than the violations of the locality assumption which were the focus of the target article.

R1.2. Is the locality assumption ever right? Many commentators saw the need to emphasize more strongly than I had done that the locality assumption may be right in some cases. I agree with this, and would like to distinguish two ways in which it is true. First, as stated in sections 1.4 and 3.3.3, there may be some cognitive domains in which the locality assumption generally holds true. Second, even in domains in which it does not generally hold true, there may be many specific instances in which it is not violated. Note, however, that, in the latter case, we would not want to say the locality assumption is true, any more than we would want to say the assumption that cats are black is true just because some cats are black. My goal was to argue that as a general assumption, the locality assumption is wrong.

For example, **Butterworth** points out that although some abnormal performances are best explained using the assumptions of PDP as opposed to the locality assumption it does not follow that all are. This is true, but not a criticism of my position. Similarly, **Young** offers counterexamples to the claim that everything interacts with everything, citing as an example the fact that language impairment does not affect toe wiggling. The claim that is up for discussion is a different one, however: that many things interact with many things, including things that are not logically necessary or intuitively relevant, such as the interaction between visual and functional semantics when functional semantic information is being accessed from a verbal query.

Grodzinsky & Hadar draw a line between domains that might be compatible with PDP, and at least one domain

that they view as surely incompatible, namely, language. They contrast this position with an inaccurate characterization of my own: "Farah's claim . . . is made as if all cognitive domains were alike (at least with respect to modularity)." As I said in the target article, it is indeed possible that some domains, such as language, will conform to different information-processing principles. On the substantive issue of whether interactive or encapsulated accounts of language are correct, and the consequent validity of the locality assumption in neurolinguistics, I differ from these commentators only in being more agnostic than they appear to be on the basis of currently available evidence.

R. Campbell emphasizes the heuristic role of the locality assumption in neuropsychology as a reason to retain it. Hypotheses about dissociations have to come from somewhere, and the locality assumption can provide a good starting point. I agree, although the problem remains that this heuristic will bias us toward noninteractionist hypotheses. **Carey & Milner** view this as an acceptable risk for neuropsychologists working on vision, because of independent evidence that the visual system has a modular organization. There is certainly abundant evidence for modularity in the sense of specialized functions (see sect. R1.1), but there is also evidence of significant interactivity, particularly among the systems involved in visual attention (e.g., the competitive phenomena alluded to by **Kinsbourne**). **McCarthy** suggests that one's a priori conceptions of cognitive or perceptual "domains," such as faces and animals, can also be a heuristic in formulating hypotheses, so that one is not a slave to the locality assumption. I agree with this too, although it is no panacea because our conceptions of domains are not guaranteed to be correct, and indeed often rely on prior locality assumption-based neuropsychology!

The truth of the locality assumption is clearly an empirical issue, but as I pointed out in the target article (sect. 1.3), it is not one that lends itself to a single critical experiment. Rather, it can be decided only by evaluating the whole body of known neuropsychological phenomena with respect to alternative explanations that involve modular and interactionist architectures. Because neuropsychologists have only recently begun to consider the latter kinds of explanations, and because we lack good heuristics for generating such explanations, we may be in a state of uncertainty for some time. The current evidence seems to bracket the likely answer somewhere between the extremes of "local damage always has local effects" and "local damage never has local effects." How often the locality assumption is right and whether there are certain domains where it is most likely to be right are important open questions. Finally, it is worth noting that progress on these questions will be closely coupled with progress on questions of interactionist versus encapsulated architectures for explaining normal cognition, questions that have occupied many of the finest minds on both sides of the issue. With respect to the present commentaries, work such as that described by **J. Campbell, Clark, and Sekuler** suggests a narrower rather than a wider range of applicability of the locality assumption. Physiological studies of normal brains are also relevant, and the evidence cited by **Posner** favors the locality assumption in the case of disengaging attention.

R1.3. Straw man characterization of neuropsychology? Although many commentators explicitly granted that the locality assumption is ubiquitous in neuropsychology, some disagreed with this generalization. **Semenza** states that "at the present time, nobody would subscribe to the locality assumption as Farah seems to intend it," **Grodzinsky & Hadar** say they "do not believe there is even one 'modularist' who would subscribe to it," and **Butterworth** can think of "no one who has maintained, in print, that all components of the brain's functional architecture are informationally encapsulated." They go on to highlight the implausibility of the locality assumption, making a compelling case that no sane neuropsychologist would hold this assumption and that the type of neuropsychology I discuss must therefore be a "straw man."

These commentators are probably right that most neuropsychologists would hesitate to endorse a bald statement of the locality assumption and the generalized informational encapsulation that is entailed by it. The relevant issue, however, concerns whether the research practices of these same neuropsychologists implicitly depend upon the locality assumption. There are undoubtedly many unexamined background assumptions in psychology and some do seem utterly wrong when stated explicitly and taken to their logical conclusions (e.g., recall the unpalatable implications of the "language of thought" hypothesis discussed by **Fodor**, including the conclusion that most concepts, including "trumpet," are innate). My claim is that the locality assumption is widely used in neuropsychology, however uncomfortable some neuropsychologists might be when confronted with an explicit statement of it, and in support of this I offered a variety of examples of research findings whose accepted interpretations hinge on the locality assumption. Let me add that these examples were chosen because they are, in my eyes and in the eyes of the field, examples of excellent and highly respectable neuropsychological research, not straw man examples.

Burton & Bruce point out that a growing number of neuropsychologists have found alternative means of interpreting their data without the locality assumption, using all or just some of the principles of PDP. Happily, this is true! Nevertheless, these cases remain very much the exception rather than the rule. It seems unfair to suggest that "in (at least) the case of covert recognition in prosopagnosia, Farah seems to have focussed selectively on an example of 'old-fashioned' cognitive neuropsychologizing." The "old-fashioned" theory to which they refer bears a 1992 publication date and resulted from the collaboration of two of the leading groups working on covert recognition (**de Haan et al. 1992**). This hardly constitutes picking on a straw man.

Carey & Milner suggest a different way in which I may have portrayed neuropsychologists as less prudent than they really are. They contend that neuropsychologists do not draw inferences about the normal cognitive architecture from dissociations, contrary to my description of cognitive neuropsychology. Rather, neuropsychologists are led to propose hypotheses on the basis of dissociations, which must then be tested with new data, neuropsychological and otherwise. The distinction here between data that suggest hypotheses and data that test them seems difficult to sustain. For example, what if a dissociation is observed after a certain hypothesis about

the cognitive architecture has already been suggested by some different data? Do we then describe our change of attitude toward that hypothesis in the light of the new dissociation as “making an inference” from the new dissociation? Surely we are at least making an inference about the likelihood that the hypothesized architecture is correct, which is really all we are ever doing, although in some cases these probabilities are higher than in others. In the above case, however, can we not also infer something about the likelihood of a given architecture on the basis of the dissociation that also suggests it? It seems clear that we can, and do. Imagine that we do not know about any of the converging evidence gathered by Milner, Goodale, and others for the architectural distinction between visual recognition and visuomotor guidance (e.g., Goodale et al. 1991) and we are offered a bet concerning whether these two functions are subserved by distinct systems. Perhaps one would not hazard a large sum on the existence of the two systems, even if the potential pay-off was great, but one would certainly be more inclined toward the bet if one knew about the dissociation shown by Goodale et al.’s (1991) case DF.

Although I do not concede **Carey & Milner’s** point about neuropsychological inference – I have just argued that we can and do draw inferences about the cognitive architecture on the basis of dissociations – I certainly grant them the importance of continued testing of our inferences or hypotheses. In addition, I agree that in some cases new tests will reveal when an inference based on the locality assumption is wrong. However, we cannot wait for all the relevant evidence that could possibly arrive between now and the end of time and then draw our conclusions. We are working “on-line,” and a strong and well-documented dissociation in even one patient is often grounds for an inference. The following question is therefore still relevant: Should we make inferences using the locality assumption?

R1.4. Speaking of straw men . . . Some criticisms were directed toward ideas that are either very distorted or extreme versions of what I presented in the target article, or are explicitly denied in the target article. For example, **Young** correctly identifies the key issue as being whether interactions modulate system behavior at what he terms the “coarse grain” level of cognitive-psychological description, but he directs his critique at my alleged failure to seek evidence on the issue, saying “the target article treats this issue as if it could be determined by theoretical fiat.” I direct the reader to section 1.3 to see that this is precisely the opposite of my position.

Servos & Olds say that my arguments “preclude much if not all neuropsychological research,” and imply that neuropsychology is “futile.” Whereas I certainly believe that abandoning the locality assumption complicates life for neuropsychologists, one of the main points of the target article is that neuropsychology can be done without the locality assumption. At the root of this difference may lie another more general difference between their reading of the target article and the intended meaning. They take the main issue to be localization of function, not identification of what the functions are; and they further adopt the following criterion for localization of function: “If a functional deficit frequently (or always) arises following damage to one brain region, and *not* when damage

occurs to other brain regions, it seems reasonable to assign a critical role to that brain region in the processes underlying that function.” This is analytically true, because of what it means to play a “critical” role, and I would not dream of arguing against it. Had I argued against it, however, I would understand why someone would think me pessimistic about the prospects for neuropsychology!

Butterworth characterizes the target article as an attack on both **Shallice** and **Caramazza**. The target article attacks the locality assumption. Although I maintain that most neuropsychologists have used the locality assumption, including myself as well as **Shallice** and **Caramazza**, neither of the latter specifically champions it. Section 1.1 of the target article reviews the relationship between certain published statements of their methodological precepts and the locality assumption.

R2. Problems with PDP

I use PDP as an alternative set of working assumptions in place of modularity in neuropsychology. To the extent that PDP yields more sensible, parsimonious accounts than modular accounts (and, eventually, confirmed predictions), this lends support to PDP as a description of human information processing and its neural implementation. However, many commentators found PDP so questionable as to be dangerous to use even in this relatively agnostic way.

R2.1. Constraining PDP. One criticism is that PDP models are too unconstrained to be explanatory (**Bullinaria, Burton & Bruce, R. Campbell, Carey & Milner, Grodzinky & Hadar, van Hezewijk & de Haan, Kinsbourne, Small**). For example, **Carey & Milner** point out that networks are powerful enough to do all kinds of things that may be remote from what the brain does and how it does it. **Van Hezewijk & de Haan** suggest that, for this reason, PDP models will have to be parasitic on common sense and prior theoretical ideas to confer substantive theoretical content to them. **Small** captures a key idea behind these worries when he likens PDP to a language for describing theories, one that is not in itself true or false. With due respect, this key idea is wrong! The principles summarized in section 1.4 are empirical claims, which could be true or false. Furthermore, any models that incorporate these principles are constrained by them. For example, a model that has distributed representations will show generalization and cross-talk, both psychologically relevant model properties, whether the modeler wants them or not. Likewise, interactivity commits the modeler to nonlocal effects of local damage.

Although PDP is not a neutral, general-purpose modeling medium but constrains models in ways that affect their ability to account for psychological phenomena, it is also only one source of constraint. In this sense, commentators such as **van Hezewijk & de Haan** are right in stating that PDP models must also incorporate ideas external to PDP, from common sense or psychology. For example, in the semantic memory model, we imposed the additional constraints that there are modality-specific forms of semantic knowledge and that vision and language each interact with semantic knowledge but do not interact directly with each other (in other words, one cannot name

something one sees without knowing what it is). This does not make PDP models, in van Hezewijk & de Haan's words, "ad hoc." Models are supposed to incorporate theoretical ideas. Indeed, incorporating theoretical ideas that have some independent motivation, as opposed to tailoring the model just to fit the data, keeps models from being ad hoc. Furthermore, it seems unfair to say that "PDP modelers just ask themselves: 'What is the next phenomenon to model?' whereas (realist) symbolic modelers would continue: 'So what are the predictions from this model and how can I test them?'" On what do the authors base this? Although the generation and testing of new predictions is less common than one might like with all types of models, there are numerous examples of PDP models that have motivated new empirical research. For example, the McClelland and Rumelhart (1981) interactive model of context effects in letter perception led to new empirical tests of the model (Rumelhart & McClelland 1982). The Seidenberg and McClelland (1989) model of word reading has motivated a series of new studies of reading in normal humans (Jared et al. 1990). PDP models and associated hypotheses are on at least equal footing with other hypotheses in neuropsychology with respect to their risk of being ad hoc. If anything, the principles mentioned in section 1.4 add additional constraints to PDP models.

Kinsbourne suggests that a specific way to increase the constraints on PDP models would be to require them to account for quantitative as well as qualitative aspects of the data. Although more specific predictions are of course desirable, this does not negate the value of achieving qualitative predictions. Furthermore, qualitative and quantitative predictions are on a continuum, and although most PDP models do not capture the precise values, distributions, and so on, of the dependent measures, they often predict some quantitative characteristics. For example, the model of semantic memory impairment predicts that more pronounced dissociations should be observable with living than with nonliving things (compare Figs. 3A and 3B) and that the dissociation between retrieving functional knowledge of living and nonliving things should be smaller in a given patient than the dissociation in other measures of knowledge (compare Figs. 3a and 4). The model of impaired attentional disengagement predicts the quantitative pattern of the four relevant means (Fig. 9). The model of covert recognition predicts not only that covert recognition will be partially preserved after damage to the visual face-recognition system but that it will remain relatively preserved at levels of damage at which overt recognition is close to or at chance (compare Fig. 12 and Figs. 13–15). Finally, it should be pointed out that this comment, too, applies equally well to non-PDP hypotheses in neuropsychology. We typically predict differences in a given direction, or patterns of means, but not precise effect sizes.

Kinsbourne also suggests that as long as the models are qualitative, they are superfluous: "The idea does all the work." It is of course true that the models are only of interest as tools for developing and testing hypotheses, but they are important tools. Some network behavior can be reliably intuited without simulations, and indeed Kinsbourne has been doing this since pre-PDP modeling days (e.g., Kinsbourne 1977). Nevertheless, not everyone for whom the ideas are relevant has developed the necessary

intuitions. In addition, intuition can handle only limited complexity, and even in simple cases it can be wrong!

Bullinaria proposes ten additional constraints on the way PDP models should be used in neuropsychology. This list raises interesting and important issues for modelers to consider, and any model that satisfied all of these constraints would certainly be admirable. However, it would be wrong to dismiss a model simply on the grounds that it fails to satisfy some of them. Models should be evaluated with respect to the question we want to answer with them. So, for example, if we have a question about the computational pressures toward division of labor in an architecture, then Bullinaria's rule 3 is relevant. If, however, our question is the more standard one in neuropsychology, specifically, Could an architecture with such-and-such components in such-and-such a configuration account for the observed dissociations? then insistence that the architecture be learned is gratuitous. Similarly, although I agree with Bullinaria, Shallice, and the many other writers who have pointed out that strong double dissociations provide the clearest evidence concerning the normal architecture, I do not agree with rule 1 that other types of dissociation or even associations are not worthy of modeling. If one wants to understand a particular system, one will work with whatever evidence is available about that system and one will simply be mindful of the alternative explanations permitted by it.

R2.2. What parts of PDP are doing the work? It is pointed out by Burton & Bruce that the different principles of PDP are to some extent separable. These commentators question the value of distributed representation, asking what explanatory work distributed representation does and whether the choice of a specific type of distributed representation or combination of distributed and local representation, rather than the general principle of distributed representation, is what makes the models work. They also complain that distributed representations are inherently difficult to interpret and that their units have no referents, lending "mystique" but not clarity to the models. I maintain that the concept of distributed representation does explanatory work and is far less mysterious than Burton & Bruce seem to think.

First, a point of clarification: the units of both local and distributed representations have interpretations. In a local representation, the referent is some item in the represented domain. In a distributed representation it is some "microfeature" (Hinton et al. 1986) of the items, which may or may not correspond to a nameable or intuitive feature. Another way of expressing this is that any unit has a meaning in the sense of some extension in the represented domain. Thus, semantic knowledge in the Farah and McClelland model is indeed distributed. It is even distributed in the model of covert face recognition, as "actor" and "politician" are features of the semantic representation of people. (Of course, the representation of occupation is local; whether a representation is local or distributed can only be defined relative to what it is representing.) Location is clearly local in the model of visual attention, but could be (and we hope eventually will be) distributed.

What explanatory work does distributed representation do? It seems clear that the distributedness of the semantic representations across visual and functional

knowledge is a key constituent of our account of impaired functional knowledge of living things. What about the covert recognition model? **Burton & Bruce**, as well as **R. Campbell and Young**, compare this model to an earlier interactionist model of covert recognition proposed by **Burton et al. (1991)**, which used local representation. The contrast is helpful in bringing out the role of distributed representation. **Burton et al.**'s model accounted for priming effects, one of the three types of covert recognition effect described in section 2.3.1. Both models account for this effect in the same way, by subthreshold activation of units downstream from the visual face-recognition system. The PDP principles that are doing the explanatory work here are graded representation and interactivity. However, the model with distributed representation can also account for the two other qualitatively different types of covert recognition effect, which the local model does not address. By allowing for *partial* damage to the visual representation of *all* faces, distributed representation makes it possible to account very naturally both for savings in relearning and for the preservation of the perceptual processing advantage for familiar faces. In fact, the use of distributed representations even allowed our model to account more fully for the priming data; the observed asymmetry between interference and facilitation emerges naturally from the influence of other parts of the distributed representation on the activation of the occupation features in our model. (For further discussion of the relations between the two models, see **Farah et al. 1993**.) In response to the question of whether these advantages derive from distributed representations in general or the specific set used in this model, no special tailoring was needed to obtain these findings; for both the semantic memory model and the covert recognition model the architectural assumptions were quite minimal and the representations themselves were generated randomly.

The fact that all three of these qualitatively different covert recognition effects were accounted for by the properties of distributed representation in a single fairly simple model supports the explanatory value of distributed representation. It also counters **Burton & Bruce's** charge that the use of PDP in neuropsychology represents punctate modeling of isolated phenomena.

Chater raises a similar question about interactivity. He points out that many of the connectionist models used in cognitive psychology and neuropsychology are feedforward models in contrast to the three recurrent models featured in the target article. Perhaps it is not interactivity per se but other aspects of PDP models that are responsible for their success. It is true that many PDP models are feedforward, although this is often due more to the feasibility of training recurrent nets than to a principled choice. Nevertheless, it is fair to ask: Does interactivity deserve any credit for simulating the phenomena of interest? In the semantic memory model, the interactions among the different parts of the representation are critical for producing an impairment in functional knowledge when visual knowledge is damaged. In the attention model the inhibitory interactions between the attention units and the recurrent connections from attention units to perception units are also critical. In the model of covert face recognition, the simulation of at least some of the phenomena depends on the network having attractor

states. Finally, the Hinton and Shallice model mentioned by **Chater** also has recurrent connections, which are important in the genesis of its error patterns.

R2.3. Levels of analysis. The general concept of levels of analysis came up in a number of commentaries, frequently with the suggestion that PDP's apparent superiority to modular accounts hinges on a confusion between levels of analysis. Two rather different senses of the phrase "levels of analysis" are used in these commentaries. In some cases (**Chater, McCarthy, Young**) it refers to the "grain size" of an account of causal mechanism. **Young** makes an analogy between the interactivity of neural information processing and cognitive functioning on the one hand and interactions between molecules in the air and global atmospheric phenomena on the other. **Young's** analogy is in many ways insightful and helpful in clarifying the crucial issue. The analogy makes clear that causal mechanisms can be described at each level of analysis, fine-grained and coarse-grained. It also implicitly poses the key question: How autonomous are the coarse-grained accounts of mechanism with respect to the fine-grained ones? According to **Young**, the answer, with respect to molecular and atmospheric phenomena, is that the coarse-grained level is autonomous. Weathermen do not have to know about people's breathing patterns because small local pressure changes become undetectable over meteorological distances. **Young's** analogy, however, serves as an argument against interactionism in neuropsychology only if we assume that the answer for one type of system and set of levels is the same for every other. This is clearly not true. Even if **Young's** analysis of atmospheric behavior is correct, it remains an open empirical question whether the interactivity that he grants is present at the fine-grained neuronal level will also affect behavior at the coarse-grained level of neuropsychological phenomena. Far from claiming to settle this issue by theoretical fiat, the clearly stated goal of the target article was to address this empirical issue by comparing mechanistic accounts of neuropsychological deficits based on PDP with accounts based on assumptions of minimal interactivity.

McCarthy makes a point similar to **Young's**, when she says "the fact that the brain is a complex biological system that shows nonlocal effects when injured does not necessarily entail that cognitive theories must mirror this interactivity. (For example, in the case of blood flow, hyperperfusion in the region of an infarct does not mean that the damaged areas of brain are contributing more to the cognitive performance of the system.)" Although facts about brain function are clearly more relevant to this empirical issue than facts about other systems such as the atmosphere, the particular counterexample chosen by **McCarthy** does not concern cerebral information processing but cerebral haemodynamics. Blood flow can be used as an index of regional cerebral information-processing activity, but no one would mistake it for a mechanism of information processing.

"Levels of analysis" was also used by **Oaksford** to refer to a different distinction, which also appears in **Mesulam's** framework for understanding attentional impairments. This is the distinction between the description of behavior (normal or pathological) and a causal mechanism to account for the behavior. **Oaksford** claims that the description of behavior, which he calls a "computational" ac-

count, is the proper subject matter for box-and-arrow models, as tempting as it may be to ascribe causality to these models. For example, he proposes a box-and-arrow model for semantic memory impairments with one lesion in the visual semantics box and a second lesion in a box labeled "functional semantics for living things." He points out that such a model "does appropriately summarise the patients' pattern of deficit at the computational level. And, of course, the computational level does not specify how a function is to be implemented. (Note that because a task is decomposed as requiring two functions to be computed, this does not mean that, ipso facto, two anatomically distinct causal mechanisms are required for their computation.)"

It would be folly to think that lesions to a cognitive mechanism would have local effects on behavior in the sense of being *local to behavior in some task*. This would require a one-to-one mapping of task-defined abilities and components of the cognitive architecture, a possibility that **Mesulam** correctly dismisses in his discussion of the mapping between behavioral and computational "planes." Rather, the locality assumption concerns the relation between lesions and their effects on nonlesioned components of a causal mechanism. The distinction can be illustrated using **Grodzinsky & Hadar's** discussion of phonological dyslexia. They state that in this case the locality assumption implies the existence of a component dedicated to reading nonwords. As described in section 1.2, the locality assumption implies the existence of a grapheme-to-phoneme translation mechanism. Nonword reading is a task-defined ability; grapheme-to-phoneme translation is a more plausible candidate for a component of the cognitive architecture, used during the reading of both words and nonwords.

Van Gelder identifies the erroneous assumption behind **Grodzinsky & Hadar's** reasoning as "the other locality assumption" – that for every task-defined ability there is some dedicated component of the cognitive architecture. Van Gelder suggests that many neuropsychologists hold this assumption as well as the first locality assumption. I think this is not quite fair. Although psychologists may sometimes have too much faith that their tasks effectively isolate and test a single underlying cognitive component, they are generally aware that the relation between tasks and components is potentially complex and as neuropsychologists they take themselves to be studying the loss of components rather than task-defined abilities.

R2.4. Biological realism of PDP. Several commentators discussed the biological realism of PDP models. Some found the glass half empty, others found it half full. **Mesulam** argues that PDP captures important features of brain functioning. **Wasserman** appears to agree, although he points out that the functioning of individual neurons may be much more complex than the units of PDP models suggest. In particular, he suggests that features of distributed representation at the network level may also be found at the individual neuron level. **Carey & Milner** also see PDP models as a step in the right direction, but they express concern over the large proportion of biologically unrealistic features of these models. As an example, they focus on the back-propagation learning algorithm (which was not, in fact, used in any of the three models of the

target article), pointing out that it is very implausible as a model of real learning in the brain. **Van Hezewijk & de Haan** make the same point. This is, of course, true, and has motivated many modelers to explore more biologically plausible learning algorithms. But even this glass is at least a quarter full: in many cases, including the first two models of the target article, one is not interested in modeling learning per se, and the so-called learning algorithm is just used to set the weights in the network so that it will perform the tasks of interest. The term "learning" has irrelevant psychological connotations in these cases and it might be less confusing to call such algorithms "weight-setting algorithms." Unless there is some systematic relation between the way the necessary weights are found and the aspects of model performance under study, which in general we have no reason to expect, it is harmless to use unrealistic learning algorithms.

In the first of their cautionary remarks, **Humphreys & Riddoch** point out that in order for components to be separately lesionable, they must be anatomically separate. They see this as incompatible with the connectivity hypothesized in the models of the target article because, in their words, "only systems that are anatomically proximal may develop functional interconnections." Although it is certainly true that short-range connections are more abundant than long-range connections, as argued in the anatomical and computational references they cite, it is also true that there are functionally important long-range connections. To mention just a few well-worked-out examples, consider Goldman-Rakic's (1987) findings on the anatomy and function of the circuits linking prefrontal cortex to parietal and temporal cortices, the cortical-hippocampal circuitry involved in learning and memory (e.g., Squire 1987), and indeed the findings from animals and humans on the circuitry of attention described by **Mesulam** and **Posner**.

Small points out that I have not availed myself of potentially important clues to the nature of patients' deficits from lesion location and etiology. In fact, with just the set of patients to whom Small refers, my colleagues and I have used neuropathological information as one source of evidence about the functional nature of the deficit (Farah et al. 1991, pp. 191–92). However, the inclusion or omission of such evidence does not seem directly relevant to the biological plausibility of PDP, nor to the issue of whether interactionist accounts are more appropriate than modular ones. Beyond this, I find little on which to disagree with Small. The brain reportedly is grey, wet, and slippery. Furthermore, most of Small's brief overviews of localization in neurology, history of PDP versus Von Neumann architectures, and so on, sound right to me. This leaves me wondering what the "missing neurobiological and computational links" are that weaken my arguments in his view.

Diederich criticizes PDP models on the grounds that they fail to incorporate many known features of brain function, such as the structure of different types of neurons, their patterns of connectivity, their sheer numbers, and so on. I would certainly not argue against striving for more biological realism in PDP models. However, it would be hasty to conclude that the kinds of models discussed in the target article are worthless or misleading because they fall short of full biological realism. It is a commonplace that models have theory-relevant and

theory-irrelevant attributes. It is also rather a cliché that science must often simplify nature in order to understand it. PDP models should be viewed as simplifications of the brain, having enough theory-relevant attributes of the brain to be informative on many questions but clearly leaving out or even contradicting many known aspects of brain function.

Among the theory-relevant aspects of PDP models are the use of distributed representations, which have been identified in numerous brain systems (e.g., Sparks et al. 1990; Young & Yamane 1992), the large number of inputs to and outputs from each unit, the modifiable connections between units, the existence of both inhibitory and excitatory connections, summation rules, bounded activations, and thresholds. PDP models allow us to find out what aspects of behavior, normal and pathological, can be explained by this set of theory-relevant attributes. Of course, some behavior may be explainable only with the incorporation of other features of neuroanatomy and neurophysiology not currently used in PDP models. This seems quite likely, and the discovery of such instances will be extremely informative with respect to the functional significance of these features of our biology. However, note that this problem does not apply to cases in which the current models perform well. In such cases, the only danger I can see associated with nonrealism is that the model's success might depend on a theory-irrelevant simplification. For example, scale is generally treated as theory-irrelevant, but it is possible that certain mechanisms will work only for small networks or small amounts of knowledge. We must be on the lookout for such cases, but we must also recognize that, barring a malevolent god, it is unlikely that the success of most models will depend critically on their unrealistic features.

R3. Discussion of specific models

R3.1. Semantic memory. Two points are made by Butterworth, one about the semantic memory model and one about scholarly credit for the ideas embodied in the model. Concerning the modality-specific semantic memory model, he states that it is not "formally different" from the category-specific alternative, in that there are two components of semantic memory with different specializations interposed between various more peripheral systems. In this sense of "formal" equivalence, he is right. But this seems a strange criterion for comparing models. The Ptolemaic and Copernican theories of planetary motion would also be equivalent by this criterion: there is one body in the middle with others orbiting around it. In both cases, the difference in the identity of the "forms" in the "formal" description makes a big difference! In the present case, one semantic memory model has category-specific components and the other has modality-specific components.

Butterworth's second point is that McClelland and I were saying nothing new in our model of semantic memory impairment. In support of this he furnishes quotations from the writings of Warrington and McCarthy (1987) and from Shallice (1988) indicating their use of the idea of distributed modality-specific representations. Readers of the target article (sect. 2.1.2) as well as the original report of the model will see that these contributions were explic-

itly acknowledged. What is new in our model, and not a part of the earlier theorizing, is the idea that *any portion of this representation needs collateral support from other portions to be accessible*. This allows the model to explain impaired access to functional knowledge of living things after damage to visual semantics.

Kinsbourne offers an alternative explanation of the phenomena based on the notion of "usual" versus "unusual" routes to naming. Although it seems to account for the data at hand, it also seems to predict the existence of subjects who are selectively impaired in functional knowledge of living things, while showing preservation for visual knowledge of all things (because visual knowledge is intact) and functional knowledge of nonliving things (because as a less "unusual" route for nonliving things, they will require less "functioning territory"). To my knowledge no such subjects have been reported.

Both McCarthy and Humphreys & Riddoch point out the existence of patients whose impairment in knowledge of living things affects only visual knowledge. This seems to present a challenge to the interactive model, as loss of visual knowledge should influence access to functional knowledge. The impairment of functional knowledge, however, will always be less pronounced than the impairment of visual knowledge when visual knowledge is damaged. For example, even when visual semantics is completely eliminated, model performance with functional semantics is only moderately impaired (Fig. 4). Therefore, at low enough levels of damage to visual semantics, a functional semantic impairment may be undetectable.

Van Hezewijk & de Haan ask whether the semantic memory model could accommodate the finding of preserved priming by unrecognized living things. Priming will be observed any time there is partial representation of an item within the to-be-primed set of units. Failure to name living things frequently occurs with partial representation of the item (for example, with enough features active to distinguish it from all but one or two of the foils). Hence the model would not just be able to accommodate such findings; at all but the most severe levels of damage, the model would *necessarily* show priming for living things that it fails to name.

Clark applauds the distributed and interactive features of the semantic memory model but expressed disappointment that it retains the distinction between visual and verbal input/output representations, on the one hand, and specifically "semantic" representations on the other. I agree that one of the appeals of viewing knowledge as distributed across multiple brain systems is that the ensemble of specific types of representations could function as "semantics." Our model is consistent with such a view if one interprets the semantic layer as containing the relatively more abstract components of the visual and motor systems, which are nevertheless part and parcel of those systems. However, we did not really attempt to do justice to this interesting idea, because our goal was to account for the finding of impaired functional knowledge about living things.

R3.2. Disengagement of attention. Posner raises a number of challenges to the Cohen, Romero, and Farah model of attentional disengagement, whose goal was to explain the "disengage deficit" without a dedicated "disengage component." Posner marshals data from split-brain and

parietally damaged patients, positron emission tomography (PET) studies of normal humans, and single-unit recording studies of monkeys, none of which our model currently accounts for. I will review each of these and discuss whether a suitably updated model could account for them without becoming grossly ad hoc.

The finding concerning visual search in split-brain patients is not in conflict with the basic idea of the model but only with an incidental feature of its implementation, specifically, the use of just two location representations. We do not wish to claim that the competitive interactions underlying the allocation of attention are confined to those between the left and right visual fields. Although only two locations were represented in the model, this is a simplification of an array of numerous units with topographically organized receptive fields. In this hypothetical model, severing the inhibitory connections between the representations of two regions of space would be expected to end cross-region attentional interactions but would leave the within-region interactions essentially unchanged. This is what was found with split-brain patients. Posner also raises the question of whether the parietal lobe controls the allocation of attention to particular *locations* in space, or whether it also controls the ability to disengage attention in order to move it in the contralateral *direction*. Although there are conflicting data about this (e.g., the Corbetta et al. 1993 study cited by Posner found evidence of location-based, not direction-based, attention), it is certainly true that the current model would not be able to account for phenomena involving directional attention. However, the basic idea of competitive interactions between different possible attentional allocations could in principle be extended to competition between directions of attentional movement.

Posner describes a recent PET study of visuospatial attention in which parietal activation was found during shifts of attention from one lateral position to another but not during static attention to the central visual field (Corbetta et al. 1993). On the face of things, this disconfirms our hypothesis that the parietal lobe subserves attention per se, rather than its disengagement. However, as Posner acknowledges, the result does not necessarily indicate the activity of a "disengager"; it could reflect the movement of attention.

Admittedly, all of these responses to the evidence reviewed by Posner involve various adjustments to the model or alternative interpretations of the data. Nevertheless, I think it is fair to say that none of the proposed adjustments or reinterpretations are outlandish, and therefore none of the findings discussed so far definitively rule out the Cohen et al. hypothesis. Alas, the findings of Robinson et al. (1991) pose a more serious challenge. He has recorded from parietal cortex in monkeys performing the Posner task, and found that some cells respond primarily when attention must be shifted from a distant location to the cell's receptive field. This is exactly how one would expect a "disengager" to act if one were recording from it. We await the full report of this conference abstract and recognize that in the light of this and possibly other results cited by Posner, our explanation of the disengage deficit may be wrong.

Finally, Posner makes two more general points that deserve comment. The first is that his own theorizing

about attention involves a distributed circuit, not a local center. This is true, and serves to raise an interesting distinction between different types of network hypotheses. One can imagine a circuit comprised of numerous encapsulated components, each computing its function on the basis of a relatively small number of inputs, and afterward sending just one discrete output to some small number of recipient components. This would be a distributed circuit, but diametrically opposed to the PDP networks described in the target article. The network theories of many neuroscientists such as Posner, Mesulam, Heilman (e.g., Heilman et al. 1985), and Goldman-Rakic (e.g., 1987) do not make clear which type of network is hypothesized. In none of these cases, however, are principles of PDP, such as interactivity or distributed and graded representation, called upon to do explanatory work. The second general point raised by Posner is that a combination of methods, including testing of clinical populations, brain imaging of normals, and when possible, animal experimentation, provides the most solid base for theorizing. I have always agreed with this, and seeing what new light can be shed upon the mechanisms of attentional disengagement in this manner reinforces my enthusiasm for converging methods.

In contrast to Posner, Kinsbourne endorses the Cohen et al. (in press) model of attentional disengagement, and traces the key idea of competitive interactions back to his own theorizing of twenty years ago. My colleagues and I acknowledge his seminal idea in the original article and are happy to do so here as well.

Humphreys & Riddoch also wish to go on record as having questioned the need to hypothesize a distinct mechanism for disengaging attention. They suggest, however, the need for some mechanism used in disengaging attention but not for orienting it.

Umiltà argues that visual attention is already known to involve interactivity, so that further demonstrations do not significantly change our conclusions about modularity and the locality assumption with respect to the visual attention system. The interactivity to which he refers is the influence of conscious volition on the allocation of attention. The interactivity hypothesized in the Cohen et al. model is the influence of attention to one stimulus on attention to other stimuli. Umiltà's point is certainly relevant to the issues of encapsulation and locality, but given that these issues are matters of degree, it is still of interest to discover other dimensions of interactivity in this system.

Umiltà also mentions evidence from the hemispatial neglect syndrome of multiple visual-attention systems, affecting different coordinate systems (viewer-centered, environment-centered, object-centered) and different stimulus types (faces, numbers, the human body). He suggests that these dissociations call for an encapsulated account of multiple attentional modules. Although this may be the correct conclusion, it seems far from irresistible on the basis of the data cited by Umiltà. For example, assuming interactivity between the representation of the visual field on the one hand and both spatial attention and object representations on the other, as proposed in Farah (1990, Ch. 6), partial degradation of particular object representations might be expected to synergize with a partial unilateral damage to the spatial-attention system to produce detectable neglect for just certain objects.

R3.3. Covert recognition. Most of the comments on covert recognition concerned the range of phenomena the model can explain. In addition to the findings already accounted for, what other features of prosopagnosia and covert recognition could be captured by the model (with modest extensions to simulate the relevant additional tasks) and what phenomena are, in principle, out of the model's reach? Davidoff & Renault raise two specific questions along these lines. First, they point out that prosopagnosics appear to be impaired in their processing of unfamiliar as well as familiar faces and that this finding is inconsistent with localist models of face recognition in which each familiar face has a local representation. They suggest that the way O'Reilly, Vecera, and I modeled face recognition with distributed face representations might be better able to account for this finding and ask whether it could be extended to demonstrate this explicitly. In fact, the model already demonstrates this, in the sense that its speed of perceptual analysis (settling time) for unfamiliar faces as well as familiar faces is slowed after damage. This is attributable to the fact that unfamiliar faces are represented in the same distributed network as familiar faces.

Davidoff & Renault also ask whether the model could be extended to account for psychophysiological measures of covert recognition involving galvanic skin response (GSR) and the P300 component of the event-related potential. Although it is not clear how such phenomena would be modeled by us, the limitation is not intrinsic to our model but to the current lack of information-processing interpretations for these psychophysiological measures. If Davidoff & Renault can give us an explicit account of how GSR and P300 emerge from network information processing (analogous to the way in which, say, naming responses are derived from network behavior in the simulations we report) then we will be in a position to test our model against the findings they cite.

Humphreys & Riddoch discuss the relation between the severity of prosopagnosia and covert recognition and the ability of the model to account for a lack of correlation. The most straightforward prediction of our model is that overt and covert recognition should be correlated, whereas the alternative model with distinct "face-recognition" and "consciousness" components does not predict a correlation. They cite an article by McNeil and Warrington (1991) as evidence against the predicted correlation. Let me address this evidence in particular, and the prediction more generally. The article in question describes three patients who were tested on various perceptual tasks, forced choice face-recognition tasks, and a savings-in-relearning task. McNeil and Warrington showed that performance on face-perception tasks and face-recognition tasks is not necessarily correlated: two cases showed relatively poor perception of faces with some degree of recognition, and one case showed better (but still impaired) face perception but no evidence of face recognition. This study does not directly address the relation between overt face recognition and covert face recognition. Even if we were to consider the forced choice face-recognition tasks to be tests of overt recognition (contrary to McNeil and Warrington's interpretation), performance on these tasks is correlated with, not dissociated from, degree of savings in relearning.

Although it is fair to say that our model of covert

recognition predicts a correlation between overt and covert recognition, the prediction is that patients with mild prosopagnosia will manifest covert recognition. The model shows that chance overt recognition can nevertheless be accompanied by some degree of covert recognition, so that one would expect to find severely prosopagnosic patients both with and without covert recognition.

Van Gulick begins by pointing out that our account of covert recognition is roughly what classical neuropsychology tells us to expect, given that the relevant dissociation is only a single dissociation. I agree with this in part, but think there is still something informative about the model above and beyond the notions of resources and differential task difficulty. For one thing, the model is more explicitly mechanistic than concepts such as "difficulty" or "resource." It is not necessarily an alternative to these concepts, but more a cashing out of them in explicit mechanistic terms. In addition, whereas it is post hoc to hypothesize differential resource needs to explain dissociated overt and covert recognition, the model shows how the nature of overt and covert recognition requires that they be differentially susceptible to network damage. Van Gulick also points out that the locality assumption does not seem relevant to the differences between various explanations of covert recognition. The locality assumption is indeed less directly relevant in the case of covert recognition than in the other two cases, as discussed in section 2.3.4.

Van Gulick, as well as Young and van Hezewijk & de Haan, all point out another aspect of face recognition that is not simulated by the model, namely, conscious awareness. It never occurred to me that this needed pointing out! The model is intended to explain the dissociation between performance on overt and covert recognition tasks. However, although the model does not simulate consciousness in the way that it simulates naming, savings in relearning, and so forth, it is nevertheless relevant to the issue of consciousness. To the extent that conscious awareness is *correlated* with overt recognition, we can view the model as telling us something about the *neural information-processing correlates* of conscious awareness. The elucidation of these correlates is probably the main contribution that empirical science can make toward the understanding of consciousness (Farah, in press). According to the model, the likelihood of a face's identity being consciously experienced is a function of the quality of the face representation. As the representation becomes more degraded, the likelihood of conscious awareness diminishes. As Van Gulick points out, this view fits well with Kinsbourne's (1988; 1993) "integrated field theory of consciousness," because degraded representations are less able to pull the rest of the network into a state that is consistent with their content. Indeed, this is exactly what we see in the model: in naming, for example, a degraded face representation is unable to pull the name units into a state consistent with it, that is, the state of representing the face's name, even though it can support performance in various covert recognition tasks.

R4. A closing thought: Science by rules and science by seat-of-the-pants

Neuropsychology is a very self-conscious science. We do not just concern ourselves with hypotheses about cogni-

tion and the brain; we also form hypotheses about how those hypotheses are formulated and how they should be tested. Clearly, I am no longer in a position to complain about this state of affairs! Yet I would like to distinguish between two approaches to metatheoretical issues in neuropsychology and suggest that one of them is counterproductive. Although some evidently interpreted my target article as an example of this approach, its broadest goal was in fact to challenge this approach.

I am sympathetic to the general practice of reflecting on what we do as neuropsychologists and attempting some analysis and justification of it. This seems useful, as well as intrinsically interesting. The approach I wish to challenge is the codification of our scientific practices into hard-and-fast rules. Familiar examples of these include "single dissociations are inherently ambiguous and therefore not to be used as evidence," and "there is no way to group patients a priori as having the same functional lesion and therefore group study designs are invalid." These rules call our attention to important considerations but it is wrong to view the consequences of these considerations in such black-and-white terms. Progress has been made on the basis of single dissociations and group studies. This alone is proof that such rules would be counterproductive if followed strictly. Good neuropsychology makes use of an unbounded set of background facts, heuristics, and intuitions in going from observed behavioral impairment to hypotheses about the functional architecture. It is much like flying by the seat of one's pants.

In criticizing the locality assumption, I wish to broaden rather than constrain the ways we think about the effects of local lesions in the functional architecture and the range of hypotheses we consider when interpreting the behavior of brain-damaged subjects. I certainly do not wish to add new rules to a Hoyle's Neuropsychology, such as "the effects of brain damage are never local to the damaged component, and therefore an X impairment can never be interpreted as evidence for an X component," or "one must construct a computational model before making a neuropsychological inference." However, thinking about highly interactive systems does not come naturally, and it is easy to overlook viable alternative hypotheses because one's intuitions do not encompass the behavior of such systems under damage. My goal is to persuade neuropsychologists to consider such alternative hypotheses, and when necessary to educate and check their intuitions with computational models.

References

Letters *a* and *r* appearing before authors' initials refer to target article and response respectively.

- Allport, D. A. (1985) Distributed memory, modular subsystems, and dysphasia. In: *Current perspectives in dysphasia*, ed. S. K. Newman & R. Epstein. Churchill Livingstone. [aMJF]
 (1989) Visual attention. In: *Foundations of cognitive science*, ed. M. I. Posner. MIT Press. [aMJF]
 Anderson, J. R. (1978) Arguments concerning representation for mental imagery. *Psychological Review* 85:249-77. [aMJF]
 Ansell, B. & Flower, C. (1982) Aphasic adults' use of heuristic and structural linguistic cues for analysis. *Brain and Language* 16:62-72. [EZ]
 Auerbach, S. H. (1986) Neuroanatomical correlates of attention and memory disorders in traumatic brain injury: An application of neurobehavioral subtypes. *Journal of Head Trauma and Rehabilitation* 1:1-12. [SLS]

- Basso, A., Capitani, E. & Laiacoma, M. (1988) Progressive language impairment without dementia: A case with isolated category specific semantic defect. *Journal of Neurology, Neurosurgery and Psychiatry* 51:1201-7. [aMJF]
 Bauer, R. M. (1984) Autonomic recognition of names and faces in prosopagnosia: A neurophysiological application of the Guilty Knowledge Test. *Neuropsychologia* 22:457-69. [JDa]
 Behrmann, M. & Bub, D. (1992) Surface dyslexia and dysgraphia: Dual routes, a single lexicon. *Cognitive Neuropsychology* 9(3):209-58. [DCP]
 Benson, D. F. & Geschwind, N. (1989) The aphasias and related disturbances. In: *Clinical neurology*, ed. R. Joynt. J. B. Lippincott. [SLS]
 Bever, T. G., Fodor, J. A. & Garrett, M. (1968) A formal limitation of associationism. In: *Verbal behavior and general behavior theory*, ed. T. R. Dixon & D. L. Horton. Prentice-Hall. [JMC]
 Biederman, I. (1990) Higher-level vision. In: *Visual cognition and action*, ed. D. N. Osherson, S. M. Kosslyn & J. M. Hollerbach. MIT Press. [aMJF]
 Bishop, D. V. M. (1992) The underlying nature of specific language impairment. *Journal of Child Psychology and Psychiatry* 33:3-66. [RC]
 Boring, E. G. (1942) *Sensation and perception in the history of experimental psychology*. Appleton-Century-Crofts. [RS]
 Broca, P. P. (1861) Nouvelle observation d'aphémie produite par une lésion de la partie postérieure des deuxième et troisième circonvolutions frontales. *Bulletin Société Anatomie (Paris)* 6:398-407. [SLS]
 Bruce, V., Burton, A. M. & Craw, I. (1992) Modelling face recognition. *Philosophical Transactions of the Royal Society* 335:121-28. [AMB]
 Bruce, V. & Young, A. W. (1986) Understanding face recognition. *British Journal of Psychology* 77:305-27. [AMB, JDa, RAM]
 Bullinaria, J. A. (1993) Neural network models of reading without Wickelfeatures. (Submitted.) [JAB]
 Bullinaria, J. A. & Chater, N. (1993) Double dissociation in artificial neural networks: Implications for neuropsychology. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*. Erlbaum. [JAB, NC]
 Burton, A. M., Bruce, V. & Johnston, R. A. (1990) Understanding face recognition with an interactive activation model. *British Journal of Psychology* 81:361-80. [AMB]
 Burton, A. M., Young, A. W., Bruce, V., Johnston, R. A. & Ellis, A. W. (1991) Understanding covert recognition. *Cognition* 39:129-66. [rMJF, AMB, RC, AWY]
 Campbell, J. I. D. (1992) In defense of the encoding-complex approach: Reply to McCloskey, Macaruso & Whetstone. In: *The nature and origins of mathematical skills*, ed. J. I. D. Campbell. Elsevier. [JIDC]
 (1993) The architecture of numerical cognition: Modular or integrated mechanisms? *International Journal of Psychology* (in press). [JIDC]
 Campbell, J. I. D. & Clark, J. M. (1988) An encoding-complex view of cognitive number processing: Comment on McCloskey, Sokol & Goodman (1986). *Journal of Experimental Psychology: General* 117:204-14. [JIDC, JMC]
 (1992) Cognitive number processing: An encoding-complex perspective. In: *The Nature and origin of mathematical skills*, ed. J. I. D. Campbell. Elsevier. [JIDC, JMC]
 Campbell, J. I. D. & Oliphant, M. (1992) Representation and retrieval of arithmetic facts: A network-interference model and simulation. In: *The nature and origin of mathematical skills*, ed. J. I. D. Campbell. Elsevier. [JMC]
 Campbell, R. (1992) Speech in the head? Rhyme skill, reading and immediate memory in the deaf. In: *Auditory imagery*, ed. D. Reisberg. Erlbaum. [RC]
 Caplan, D. (1981) On the cerebral localization of linguistic functions: Logical and empirical issues surrounding deficit analysis and functional localization. *Brain and Language* 14:120-37. [aMJF]
 Caplan, D. & Futter, C. (1986) Assignment of thematic roles by an agrammatic aphasic patient. *Brain and Language* 27:117-35. [EZ]
 Caramazza, A. (1984) The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language* 21:9-20. [aMJF, CS]
 (1986) On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition* 5:41-66. [aMJF, BB, DCP]
 (1992) Is cognitive neuropsychology possible? *Journal of Cognitive Neuroscience* 4:80-95. [aMJF, DCP, CS]
 Caramazza, A. & Hillis, A. E. (1990) Spatial representation of words in the brain implied by studies of a unilateral neglect patient. *Nature* 346:267-69. [CU]
 Caramazza, A., Hillis, A. E., Rapp, B. C. & Romani, C. (1990) The multiple

- semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology* 7:161-90. [aMJF, RAM]
- Caramazza, A. & Zurif, E. B. (1976) Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language* 3:572-82. [EZ]
- Charcot, J. M. (1883) Un cas de suppression brusque et isolée de la vision mentale des signes et objets (formes et couleurs). *Progrès Medical* 11:569-71. [JDa]
- Chater, N. & Oaksford, M. (1990) Autonomy, implementation and cognitive architecture: A reply to Fodor and Pylyshyn. *Cognition* 34:93-107. [MO]
- Clark, J. M. & Campbell, J. I. D. (1991) Integrated versus modular theories of number skills and acalculia. *Brain and Cognition* 17:204-39. [JIDC, JMC]
- Cohen, J. D., Romero, R. D. & Farah, M. J. (in press) Disengaging from the disengage function: The relation of macrostructure to microstructure in parietal attentional deficits. *Journal of Cognitive Neuroscience*. [arMJF, GWH, DCP, CU]
- Cohen, L. & Dehaene, S. (1991) Neglect dyslexia for numbers? A case report. *Cognitive Neuropsychology* 8:39-58. [CU]
- Coltheart, M. (1984) Editorial. *Cognitive Neuropsychology* 1:1-8. [GWH]
- (1985) Cognitive neuropsychology and the study of reading. In: *Attention and performance XI*, ed. M. I. Posner & O. S. M. Marin. Erlbaum. [aMJF]
- Corbetta, M., Miezin, F. M., Shulman, G. L. & Peterson, S. E. (1993) A PET study of visuospatial attention. *Journal of Neuroscience* 13:1202-26. [rMJF, MIP]
- Covey, A. (1985) Aspects of cortical organization related to selective attention and selective impairments of visual perception. In: *Attention and performance XI*, ed. M. I. Posner & O. S. M. Marin. Erlbaum. [GWH]
- Crick, F. (1989) The recent excitement about neural networks. *Nature* 337:129-32. [DPC]
- Daffner, K., Ahern, G., Weintraub, S. & Mesulam, M.-M. (1990) Dissociated neglect behavior following sequential strokes to the right hemisphere. *Annals of Neurology* 28:97-101. [M-MM]
- Damasio, A. R. & van Hoesen, G. W. (1985) The limbic system and the localization of herpes simplex encephalitis. *Journal of Neurology, Neurosurgery, and Psychiatry* 48:297-301. [SLS]
- Damasio, H. & Damasio, A. R. (1989) *Lesion analysis in neuropsychology*. Oxford University Press. [SLS]
- Davidoff, J. & Landis, T. (1990) Recognition of unfamiliar faces in prosopagnosia. *Neuropsychologia* 28:1143-61. [JDa]
- Davis, S. N., Lester, R. A., Reymann, K. G. & Collingridge, G. L. (1989) Temporarily distinct pre- and post-synaptic mechanisms maintain long-term potentiation. *Nature* 338:283-321. [RvH]
- de Haan, E. H. F., Bauer, R. M. & Creve, K. W. (1992) Behavioural and physiological evidence for covert face recognition in a prosopagnosic patient. *Cortex* 28:77-95. [arMJF, JDa, DCP, RVG, RvH]
- de Haan, E. H. F., Young, A. & Newcombe, F. (1987a) Faces interfere with name classification in a prosopagnosic patient. *Cortex* 23:309-16. [aMJF]
- (1987b) Face recognition without awareness. *Cognitive Neuropsychology* 4:385-416. [aMF]
- Dennett, D. C. & Kinsbourne, M. (1992) *Time and the observer: The where and when of consciousness in the brain*. *Behavioral and Brain Sciences* 15:183-247. [RVG]
- DeRenzi, E. (1982) *Disorders of space exploration and cognition*. Wiley. [CU]
- (1986) Current issues in prosopagnosia. In: *Aspects of face processing*, ed. H. D. Ellis, M. A. Jeeves, F. Newcombe & A. Young. Martinus Nijhoff. [aMJF]
- DeRenzi, E., Gentilini, M. & Barbieri, C. (1989) Auditory neglect. *Journal of Neurology, Neurosurgery, and Psychiatry* 52:613-17. [CU]
- Dixon, N. F. (1981) *Preconscious processing*. Wiley. [JDa]
- Dodd, B. & Murphy, J. (1992) Language without sound: Prelingual deafness in development. In: *Mental lives: Case studies in cognition*, ed. R. Campbell. Blackwell. [RC]
- Downing, C. J. & Pinker, S. (1985) The spatial structure of visual attention. In: *Attention and performance XI*, ed. M. I. Posner & O. S. M. Marin. Erlbaum. [CU]
- Driver, J. & Halligan, P. W. (1991) Can visual neglect operate in object-centered co-ordinates? An affirmative single-case study. *Cognitive Neuropsychology* 8:475-96. [CU]
- Dunn, J. C. & Kirsner, K. (1988) Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review* 95:91-101. [JAB]
- Eagleson, R. & Carey, D. P. (1992) Connectionist networks do not model brain function. *Behavioral and Brain Sciences* 15:734-35. [DPC]
- Ellis, A. & Young, A. W. (1988) *Human cognitive neuropsychology*. Erlbaum. [RAM]
- Farah, M. J. (1990) *Visual agnosia: Disorders of object recognition and what they tell us about normal vision*. Bradford Books/MIT Press. [rMJF]
- (1991) Patterns of co-occurrence among the associative agnosias: Implications for visual object representation. *Cognitive Neuropsychology* 8:1-19. [RC]
- (in press) Visual perception and visual awareness after brain damage: A tutorial overview. In: *Conscious and unconscious information processing: Attention and performance XIV*, ed. M. Moscovitch & C. Umiltà. MIT Press. [rMJF]
- Farah, M. J. & McClelland, J. L. (1991) A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* 120(4):339-57. [arMJF, BB, GWH, RAM, DCP, RvH]
- Farah, M. J., McMullen, P. A. & Meyer, M. M. (1991) Can recognition of living things be selectively impaired? *Neuropsychologia* 29:185-93. [arMJF, RvH]
- Farah, M. J., O'Reilly, R. C. & Vecera, S. P. (1993) Dissociated overt and covert recognition as an emergent property of lesioned neural networks. *Psychological Review* 100:571-88. [arMJF, RC, GWH, DCP, RvH]
- Feldman, J. A. & Ballard, D. (1982) Connectionist models and their properties. *Cognitive Science* 6:205-54. [SLS]
- Feldman, J. A., Fianty, M. A. & Goddard, N. (1988) Computing with structured neural networks. *IEEE Computer* 21(3):91-103. [JDI]
- Felleman, D. J. & Van Essen, D. C. (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1:1-47. [DPC, JDI, RS]
- Ferrier, D. (1886) *The functions of the brain*. Smith, Elder. [aMJF, JDa]
- Fiorelli, M., Blin, J., Bakchine, S., Laplane, D. & Baron, J. C. (1991) PET studies of cortical diaschisis in patients with motor hemi-neglect. *Journal of Neurological Science* 104:135-42. [M-MM]
- Flourens, P. (1846) *Phrenology examined*, 2nd ed. (trans. Charles De Lucena Meigs). Hogan & Thompson. [TvG]
- Fodor, J. A. (1983) *The modularity of mind*. Bradford Books/MIT Press. [arMJF, NC, RAM, DCP, CU, TvG, RvH, EZ]
- Fodor, J. A., Garrett, M. & Swinney, D. (1992) A modular effect in parsing. Unpublished manuscript, Cognitive Science Department, Rutgers University. [EZ]
- Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3-71. [DPC, MO]
- Forster, K. (1979) Levels of processing and the structure of the language processor. In: *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*, ed. W. Cooper & E. Walker. Erlbaum. [EZ]
- Foster, C. L. (1992) *Algorithms, abstraction and implementation*. Academic Press. [MO]
- Freud, S. (1891/1953) *On aphasia*, trans. E. Stengel. Imago. [JDa, SLS]
- Fuhrer, M. J. & Eriksen, C. W. (1960) The unconscious perception of the meaning of verbal stimuli. *Journal of Abnormal and Social Psychology* 61:432-39. [JDa]
- Gainotti, G., Messerli, P. & Tissot, R. (1972) Quantitative analysis of unilateral spatial neglect in relation to lateralisation of cerebral lesions. *Journal of Neurology, Neurosurgery, and Psychiatry* 35:545-50. [CU]
- Garrett, M. (1981) Objects of psycholinguistic inquiry. *Cognition* 10:97-101. [EZ]
- Geschwind, N. (1985) Mechanisms of change after brain lesions. *Annals of the New York Academy of Sciences* 457:1-11. [JAB]
- Gluck, M. A. & Rumelhart, D. E. (1990) *Neuroscience and connectionist theory*. Erlbaum. [PS]
- Goldberg, M. E. & Segraves, M. A. (1987) Visuospatial and motor attention in the monkey. *Neuropsychologia* 25:107-18. [M-MM]
- Goldman-Rakic, P. S. (1987) Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In: *Handbook of physiology: Nervous system*. Vol. 5: *Higher functions of the brain*, ed. F. Plum. American Psychological Society. [rMJF]
- Goodale, M. A. & Milner, A. D. (1992) Separate visual pathways for perception and action. *Trends in Neuroscience* 15:20-25. [DPC]
- Goodale, M. A., Milner, A. D., Jakobson, L. S. & Carey, D. P. (1991) A neurological dissociation between perceiving objects and grasping them. *Nature* 349:154-56. [rMJF, DPC]
- Green, J., Morris, J. C., Sandson, J., McKeel, D. W. & Miller, J. W. (1990) Progressive aphasia: A precursor of global dementia? *Neurology* 40:423-29. [SLS]
- Gregory, R. L. (1961) The brain as an engineering problem. In: *Current problems in animal behaviour*, ed. W. H. Thorpe & O. L. Zangwill. Cambridge University Press. [RAM]
- Grodzinsky, Y. (1986) Language deficits and the theory of syntax. *Brain and Language* 27:135-59. [EZ]
- (1990) *Theoretical perspectives on language deficits*. MIT Press. [YG]

References/Farah: Neuropsychological inference

- Grossberg, S. (1987) Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11:23-63. [RvH]
- Guariglia, C. & Antonucci, G. (1992) Personal and extrapersonal space: A case of neglect dissociation. *Neuropsychologia* 30:1001-9. [CU]
- Halligan, P. W. & Marshall, J. C. (1991) Left neglect for near but not far space in man. *Nature* 350:498-500. [CU]
- (1992) Left visuo-spatial neglect: A meaningless entity? *Cortex* 28:525-35. [CU]
- Hart, G. & Gordon, B. (1992) Neural subsystems for object knowledge. *Nature* 359:60-64. [RAM]
- Hart, J., Berndt, R. S. & Caramazza, A. (1985) Category specific naming deficit following cerebral infarction. *Nature (London)* 316:439-40. [BB]
- Hatfield, F. M., Barber, J., Jones, C. & Morton, J. (1977) Object naming in aphasia: The lack of an effect of context or realism. *Neuropsychologia* 15:717-27. [RAM]
- Hay, D. C. & Young, A. W. (1982) The human face. In: *Normality and pathology in cognitive functions*, ed. A. W. Ellis. Academic Press. [AMB]
- Heilman, K. M., Watson, R. T. & Valenstein, E. (1985) Neglect and related disorders. In: *Clinical neuropsychology*, 2nd ed., ed. K. M. Heilman & E. Valenstein. Oxford University Press. [rMJF]
- Henderson, J. M. & Macquistan, A. D. (1993) The spatial distribution of attention following an exogenous cue. *Perception & Psychophysics* 53:221-30. [CU]
- Hildreth, E. C. & Ullman, S. (1989) The computational study of vision. In: *Foundations of cognitive science*, ed. M. I. Posner. MIT Press. [aMJF]
- Hillis, A. E. & Caramazza, A. (1991) Category-specific naming and comprehension impairment: A double dissociation. *Brain* 114:2081-94. [aMJF]
- Hinton, G. E. (1992) How neural networks learn from experience. *Scientific American* 267:105-9. [DPC]
- Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986) Distributed representations. In: *Parallel distributed processing: Explorations in the microstructure of cognition*, ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [arMJF]
- Hinton, G. E. & Shallice, T. (1991) Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review* 98(1):74-95. [aMJF, AMB, NC, DCP]
- Homa, D., Haver, B. & Schwartz, T. (1976) Perceptibility of schematic face stimuli: Evidence for a perceptual Gestalt. *Memory and Cognition* 4:176-85. [JDa]
- Humphreys, G. W. & Bruce, V. (1989) *Visual cognition: Computational, experimental and neuropsychological perspectives*. Erlbaum. [aMJF]
- Humphreys, G. W. & Evett, L. J. (1985) Are there independent lexical and nonlexical routes in word processing? An evaluation of dual-route theory of reading. *Behavioral and Brain Sciences* 8:689-740. [JAB]
- Humphreys, G. W., Freeman, T. & Muller, H. J. (1992) Lesioning a connectionist model of visual search: Selective effects of distractor grouping. *Canadian Journal of Psychology* 46:417-60. [aMJF, AMB]
- Humphreys, G. W. & Riddoch, M. J. (1987) *Visual object processing: A cognitive neuropsychological approach*. Erlbaum. [aMJF]
- (1993) Interactions between object- and space-vision revealed through neuropsychology. In: *Attention and performance XIV*, ed. D. E. Meyer & S. Kornblum. MIT Press. [aMJF, GWH]
- Jackendoff, J. A. (1987) *Consciousness and the computational mind*. Cambridge University Press. [RvH]
- Jackson, J. H. (1873) On the anatomical and physiological localization of movements in the brain. *Lancet* 1:84-85, 162-64, 232-34. [aMJF]
- (1878) On affections of speech from diseases of the brain. *Brain* 1:304-30. [SLS]
- Jacobs, R. A. & Jordan, M. I. (1992) Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience* 4:323-36. [DPC, GWH]
- Jakobson, L. S., Archibald, Y. M., Carey, D. P. & Goodale, M. A. (1991) A kinematic analysis of reaching and grasping movements in a patient recovering from optic ataxia. *Neuropsychologia* 29:803-9. [DPC]
- Jared, D., McRae, K. & Seidenberg, M. (1990) The basis of consistency effects in word naming. *Journal of Memory & Language* 29:687-715. [rMJF]
- Kelley, H. H. (1992) Common-sense psychology and scientific psychology. *Annual Review of Psychology* 43:1-25. [RS]
- Kertesz, A., ed. (1983) *Localization in neuropsychology*. Academic Press. [SLS]
- Kettner, R., Marcario, J. & Port, N. (1993) A neural network model of cortical activity during reaching. *Journal of Cognitive Neuroscience* 5:14-33. [DPC]
- Kimberg, D. Y. & Farah, M. J. (in press) A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General*. [aMJF]
- Kinsbourne, M. (1970) A model for the mechanism of unilateral neglect of space. *Transactions of the American Neurological Association* 95:143-46. [MK]
- (1971) Cognitive deficit: Experimental analysis. In: *Psychobiology*, ed. J. L. McCaugh. Academic Press. [aMJF]
- (1977) Hemi-neglect and hemispheric rivalry. In: *Advances in neurology*, ed. E. A. Weinstein & R. P. Friedland. Raven Press. [arMJF]
- (1987) Mechanisms of unilateral neglect. In: *Neurophysiological and neuropsychological aspects of spatial neglect*, ed. M. Jeannerod. Elsevier. [MK]
- (1988) Integrated field theory of consciousness. In: *Concept of consciousness in contemporary science*, ed. A. J. Marcel & E. Bisiach. Oxford University Press. [rMJF, MK, RVG]
- (1993) Integrated cortical field model of consciousness. In: *Ciba Foundation Symposium 174: Experimental and Theoretical Studies of Consciousness*, ed. G. R. Bock & J. Marsh. Wiley. [rMJF, RVG]
- (in press) Orientational bias model of unilateral neglect: Evidence from attentional gradients within hemispace. In: *Unilateral neglect: Clinical and experimental studies*. ed. I. H. Robertson & J. C. Marshall. Erlbaum. [MK]
- Klahr, D., Langley, P. & Neches, R. (1987). *Production system models of learning and development*. MIT Press. [aMJF]
- Kosslyn, S. M., Flynn, R. A., Amsterdam, J. B. & Wang, G. (1990) Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition* 32:203-77. [aMJF]
- Kosslyn, S. M. & Van Kleeck, M. (1990) Broken brains and normal minds: Why Humpty Dumpty needs a skeleton. In: *Computational neuroscience*, ed. E. Schwartz. MIT Press. [aMJF]
- Kunst-Wilson, W. R. & Zajonc, R. B. (1980) Affective discrimination of stimuli that cannot be recognized. *Science* 207:557-58. [RS]
- LaBerge, D. (1990) Thalamic and cortical mechanisms of attention suggested by recent positron emission tomographic experiments. *Journal of Cognitive Neuroscience* 2:358-72. [MIP]
- Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge*, ed. I. Lakatos & A. Musgrave. Cambridge University Press. [RvH]
- Lashley, K. S. (1950) In search of the engram. In: *Symposia for the Society for Experimental Biology, Number 4*. Cambridge University Press. [SLS]
- Levin, B. & Pinker, S., eds. (1991) Lexical and conceptual semantics. *Cognition* 41: Special issue. [EZ]
- Luck, S. J., Hillyard, S. A., Mangun, G. R. & Gazzaniga, M. S. (1989) Independent hemispheric attentional systems mediate visual search in split-brain patients. *Nature* 342:543-45. [MIP]
- Lynch, J. C., Mountcastle, V. B., Talbot, W. H., & Yin, T. C. T. (1977) Parietal lobe mechanisms for directed visual attention. *Journal of Neurophysiology* 40:362-89. [M-MM]
- Mandler, G., Nakamura, Y. & Shebo van Zandt, B. J. (1987) Nonspecific effects of exposure on stimuli that cannot be recognized. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:646-48. [RS]
- Marie, P. (1906) The third left frontal convolution plays no special role in the function of language. In: *Pierre Marie's papers on speech disorders*, trans. Cole & Cole (1971). Hafner Press. [SLS]
- Marin, O. S., Saffran, E. M. & Schwartz, M. (1976) Dissociation of language in aphasia: Implications for normal functions. *Annals of the New York Academy of Sciences* 280:868-84. [CS]
- Mark, V. W., Kooistra, C. A. & Heilman, K. M. (1988) Hemispatial neglect is affected by non-neglected stimuli. *Neurology* 38:1207-11. [CU]
- Marr, D. (1982) *Vision*. Freeman. [RvH, MO]
- Marshall, J. C. & Newcombe, F. (1966) Syntactic and semantic errors in paralexia. *Neuropsychologia* 4:169-76. [AWY]
- Massaro, D. (1988) Some criticisms of connectionist models of human performance. *Journal of Memory and Language* 27:213-34. [RC]
- Mazzoni, P., Anderson, R. A. & Jordan, M. I. (1991a) A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences of the United States of America* 88:4433-37. [DPC]
- (1991b) A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a. *Cerebral Cortex* 1:293-307. [DPC]
- McCarthy, R. A. & Warrington, E. K. (1986) Phonological reading: Problems and paradoxes. *Cortex* 22:359-80. [RAM]
- (1990) *Cognitive neuropsychology: A clinical introduction*. Academic Press. [RAM, AWY]
- McClelland, J. L. (1981) Retrieving general and specific information from

- stored knowledge of specifics. *Proceedings of the Third Annual Meeting of the Cognitive Science Society*:170–72. [AMB]
- (1991) Toward a theory of information processing in graded random, interactive networks. *Technical Report PDP.CNS.91.1*, Department of Psychology, Carnegie Mellon University. [NC]
- McClelland, J. L. & Elman, J. L. (1986) Interactive processes in speech perception: The TRACE model. In: *Parallel distributed processing*, vol. 2, ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [NC]
- McClelland, J. L. & Rumelhart, D. E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88:375–407. [rMJF]
- (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review* 89:60–94. [rMJF]
- McCloskey, M. (1991) Networks and theories: The place of connectionism in cognitive science. *Psychological Science* 2:387–95. [rMJF]
- (1992) Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition* 44:107–57. [JIDC]
- McCloskey, M., Caramazza, A. & Basili, A. G. (1985) Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brain and Cognition* 4:171–96. [JIDC]
- McCloskey, M., Macaruso, P. & Whetstone, T. (1992) The functional architecture of numerical processing mechanisms: Defending the modular model. In: *The nature and origins of mathematical skills*, ed. J. I. D. Campbell. Elsevier. [JIDC, JMC]
- McCloskey, M., Sokol, S. M. & Goodman, R. A. (1986) Cognitive processes in verbal-number production: Inferences from the performance of brain-damaged subjects. *Journal of Experimental Psychology: General* 115:307–30. [JIDC, JMC]
- McNeil, J. E. & Warrington, E. K. (1991) Prosopagnosia: A reclassification. *Quarterly Journal of Experimental Psychology* 43A:267–88. [rMJF, GWH]
- Merigan, W. H. & Maunsell, J. H. R. (1993) How parallel are the primate visual pathways? *Annual Review of Neuroscience* 16:396–402. [DPC]
- Mesulam, M.-M. (1981) A cortical network for directed attention and unilateral neglect. *Annals of Neurology* 10:309–25. [M-MM]
- (1990) Large-scale neurocognitive networks and distributed processing for attention, language and memory. *Annals of Neurology* 28:597–613. [M-MM]
- Miller, G. & Fellbaum, C. (1991) Semantic networks of English. *Cognition* 41:197–229. [EZ]
- Milner, A. D. & Goodale, M. A. (1993) Visual pathways to perception and action. In: *The visually responsive neuron: From basic neurophysiology to behavior (Progress in Brain Research, vol. 95)*, ed. T. P. Hicks, S. Molotchnikoff & T. Ono. Elsevier. [DPC]
- Milner, A. D., Perrett, D. I., Johnston, R. S., Benson, P. J., Jordan, T. R., Heeley, D. W., Bettucci, D., Mortara, F., Mutani, R., Terazzi, E. & Davidson, D. L. W. (1991) Perception and action in visual form agnosia. *Brain* 114:405–28. [DPC]
- Morcraft, R. J., Geula, C. & Mesulam, M.-M. (1993) Architecture of connectivity within a cingulo-fronto-parietal neurocognitive network for directed attention. *Archives of Neurology* 50:279–84. [M-MM]
- Moscovitch, M. & Umiltà, C. (1990) Modularity and neuropsychology: Modules and central processes in attention and memory. In: *Modular deficits in Alzheimer-type dementia*, ed. M. F. Schwartz. MIT Press. [aMJF, CU]
- (1991) Conscious and nonconscious aspects of memory: A neuropsychological framework of modules and central systems. In: *Perspectives on cognitive neuroscience*, ed. R. G. Lister & H. J. Weingartner. Oxford University Press. [CU]
- Movellan, J. R. (1990) Contrastive Hebbian learning in the continuous Hopfield model. In: *Proceedings of the 1989 Connectionist Models Summer School*, ed. D. S. Touretzky, G. E. Hinton & T. J. Sejnowski. Morgan Kaufmann. [aMJF]
- Mozer, M. C. & Behrmann, M. (1990) On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia. *Journal of Cognitive Neuroscience* 2(2):96–123. [aMJF, DCP]
- Nebes, R. D. (1989) Semantic memory in Alzheimer's disease. *Psychological Bulletin* 106:377–94. [aMJF]
- Newcombe, F., Mehta, Z. & de Haan, E. F. (in press) Category-specificity in visual recognition. In: *The neural bases of high-level vision: Collected tutorial essays*, ed. M. J. Farah & G. Ratcliff. Erlbaum. [aMJF]
- Norris, D. G. (1990) A dynamic-net model of human speech recognition. In: *Cognitive models of speech processing: Psycholinguistic and cognitive perspectives*, ed. C. Altmann. MIT Press. [NC]
- Oaksford, M. & Chater, N. (1991) Against logical cognitive science. *Mind & Language* 6:1–38. [MO]
- Ogden, J. A. (1987) The "neglected" left hemisphere and its contribution to visuospatial neglect. In: *Neurophysiological and neuropsychological aspects of spatial neglect*, ed. M. Jeannerod. North-Holland. [CU]
- Paivio, A. (1986) *Mental representations: A dual-coding approach*. Oxford University Press. [JMC]
- Patterson, K. E., Seidenberg, M. S. & McClelland, J. L. (1989) Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In: *Parallel distributed processing: Implications for psychology and neurobiology*, ed. R. G. M. Morris. Oxford University Press. [aMJF, AMB, NC, RAM, DCP]
- Perenin, M.-T. & Vighetto, A. (1988) Optic ataxia: A specific disruption in visuomotor mechanisms. I. Different aspects of the deficit in reaching for objects. *Brain* 111:643–74. [DPC]
- Pietrini, V., Nertimpi, T., Vaglia, A., Revello, M. G., Pinna, V. & Ferro-Milone, F. (1988) Recovery from herpes simplex encephalitis: Selective impairment of specific semantic categories with neuroradiological correlation. *Journal of Neurology, Neurosurgery, and Psychiatry* 51:1284–93. [aMJF]
- Pinker, S. (1985) Visual cognition: An introduction. In: *Visual cognition*, ed. S. Pinker. MIT Press. [aMJF]
- Plaut, D. C. & Shallice, T. (1993) Pervasive and semantic influences on visual object naming errors in optic aphasia: A connectionist account. *Journal of Cognitive Neuroscience* 5(1):89–117. [DPC, DCP]
- (in press) Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*. [DCP]
- Popper, K. R. (1963) *Conjectures and refutations: The growth of scientific knowledge*. Routledge & Kegan Paul. [RvH]
- Posner, M. I. (1978) *Chronometric explorations of mind*. Erlbaum. [aMJF]
- (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology* 32:3–25. [CU]
- (1988) Structures and functions of selective attention. In: *Master lectures in clinical neuropsychology and brain function, research, measurement, and practice*, ed. T. Boll & B. Bryant. American Psychological Association. [MIP]
- Posner, M. I. & Rothbart, M. K. (in press) Constructing neuronal theories of mind. In: *High level neuronal theories of the brain*, ed. C. Koch & J. Davis. MIT Press. [MIP]
- Posner, M. I., Walker, J. A., Friedrich, F. J. & Rafal, R. D. (1984) Effects of parietal lobe injury on covert orienting of visual attention. *Journal of Neuroscience* 4:1863–74. [aMJF, DPC, MK, CU]
- (1987) How do the parietal lobes direct covert attention? *Neuropsychologia* 25:135–45. [DPC]
- Prather, P., Shapiro, L., Zurif, E. & Swinney, D. (1991) Real-time examination of lexical processing in aphasics. *Journal of Psycholinguistic Research* (Special issue on sentence processing) 20:271–81. [EZ]
- Pribram, K. H. (1971) *Languages of the brain*. Prentice-Hall. [MK]
- Price, R. W. (1986) Neurobiology of human herpes virus infections. *CRC Critical Reviews in Clinical Neurobiology* 2:61–123. [SLS]
- Raichle, M. E. (1989) Developing a functional anatomy of the human brain with positron emission tomography. *Current Neurology* 9:161–78. [SLS]
- Reeke, G. N., Jr. & Sporns, O. (1993) Behaviorally based modelling and computational approaches to neuroscience. *Annual Review of Neuroscience* 16:507–623. [DPC]
- Renault, B., Signoret, J. L., Debrulle, B., Breton, F. & Bolgert, F. (1989) Brain potentials reveal covert face recognition in prosopagnosia. *Neuropsychologia* 27:905–12. [JDa]
- Riddoch, M. J. & Humphreys, G. W. (1987) Visual object processing in optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology* 4:131–85. [GWH]
- (1992) The smiling giraffe: An illustration of a visual memory disorder. In: *Mental lives*, ed. R. Campbell. Blackwell. [GWH]
- Robinson, D. L., Bowman, E. M. & Kertzman, C. (1991) Covert orienting of attention in macaque II: A signal in parietal cortex to disengage attention. *Society of Neuroscience Abstracts* 17:442. [rMJF, MIP]
- Rosenberg, C. R. & Sejnowski, T. K. (1986) NETalk: A parallel network that learns to read aloud. *EE & CS Technical Report #JHU-EECS-86/01*. Johns Hopkins University Press. [aMJF]
- Rumelhart, D. E., Hinton, G. E. & McClelland, J. L. (1986) A general framework for parallel distributed processing. In: *Parallel distributed processing: Explorations in the microstructure of cognition*, ed. D. E. Rumelhart & J. L. McClelland. MIT Press. [aMJF]
- Rumelhart, D. E. & McClelland, J. L. (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review* 89:60–94. [rMJF]
- (1985) Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114:193–97. [MO]
- (1986) *Parallel Distributed Processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*. MIT Press. [aMJF, PS]

References/Farah: Neuropsychological inference

- Saffran, E. M. (1982) Neuropsychological approaches to the study of language. *British Journal of Psychology* 73:317–37. [CS]
- Sartori, G. & Job, R. (1988) The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology* 5:105–32. [aMJF]
- Schott, B., Jeannerod, M. & Zahin, M. Z. (1966) L'agnosie spatiale unilatérale: Perturbation en secteur des mécanismes d'exploration et de fixation du regard. *Journal Médical de Lyon* 47:169–95. [CU]
- Schweich, M. & Bruyer, R. (in press) Heterogeneity in the cognitive manifestations of prosopagnosia: The study of a group of single cases. *Cognitive Neuropsychology*. [JDa]
- Searle, J. R. (1984) *Minds, brains and science: The 1984 Reith lectures*. British Broadcasting Corporation. [AWY]
- (1992) *The rediscovery of the mind*. MIT Press. [AWY]
- Seidenberg, M. S. (1988) Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology* 5:403–26. [GWH]
- Seidenberg, M. S. & McClelland, J. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96:323–68. [rMJF]
- Semenza, C. (1993) Methodological issues. In: *A dictionary of neuropsychology*, ed. G. Beaumont & J. Sergent. Basil Blackwell. [CS]
- Semenza, C., Bisiacchi, P. S. & Rosenthal, V. (1988) A function for cognitive neuropsychology. In: *Perspectives on cognitive neuropsychology*, ed. G. Denes, C. Semenza & P. S. Bisiacchi. Erlbaum. [CS]
- Sergent, J. (1987) Information processing and laterality effects for object and face perception. In: *Visual object processing: A cognitive neuropsychological approach*, ed. G. W. Humphreys & M. J. Riddoch. Erlbaum. [GWH]
- Shallice, T. (1988) *From neuropsychology to mental structure*. Cambridge University Press. [arMJF, BB, JAB, GWH, RAM, AWY, RVG]
- (1991) *Précis of From neuropsychology to mental structure*. *Behavioral and Brain Sciences* 14:429–69. [CS]
- Shallice, T. & Saffran, S. (1986) Lexical processing in the absence of explicit word identification: Evidence from a letter-by-letter reader. *Cognitive Neuropsychology* 3:429–58. [GWH]
- Shankweiler, D. & Crain, S. (1986) Language mechanisms and reading disorders: A modular approach. *Cognition* 24:121–37. [YG]
- Shepherd, G. M. (1990) The significance of real neuron architectures for neural network simulations. In: *Computational neuroscience*, ed. E. L. Schwartz. MIT Press. [JDi]
- Sheridan, J. & Humphreys, G. W. (1993) A verbal-semantic category-specific deficit. *Cognitive Neuropsychology* 10:143–84. [GWH]
- Shillcock, R., Lindsey, G., Levy, J. & Chater, N. (1992) A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Erlbaum. [NC]
- Silveri, M. C. & Gianotti, G. (1988) Interaction between vision and language in category-specific semantic impairment. *Cognitive Neuropsychology* 5:677–709. [aMJF]
- Smith, N. & Tsimpli, I.-M. (1991) Linguistic modularity? A case of a savant-linguist. *Lingua* 84:315–51. [RC]
- Sokol, S. M., Goodman-Schulman, R. & McCloskey, M. (1989) In defense of a modular architecture for the number processing system: Reply to Campbell & Clark. *Journal of Experimental Psychology: General* 118(1):105–10. [JMC]
- Sokol, S. M., McCloskey, M., Cohen, N. J. & Aliminos, D. (1991) Cognitive representations and processes in arithmetic: Inferences from the performance of brain-damaged patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17:355–76. [JIDC]
- Sparks, D. L., Lee, C. & Rohrer, W. H. (1990) Population coding of the direction, amplitude, and velocity of saccadic eye movements by neurons in the superior colliculus. *Cold Spring Harbor Symposia on Quantitative Biology* 55:805–11. [rMJF]
- Spurzheim, J. G. (1815) *The physiognomical system of Drs. Gall and Spurzheim*. Baldwin, Craddock & Joy. [TvG]
- Squire, L. R. (1987) *Memory and brain*. Oxford University Press. [rMJF]
- (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99:195–231. [aMJF]
- Swinney, D. (1979) Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18:645–59. [EZ]
- Swinney, D. & Fodor, J. D., eds. (1989) *Journal of Psycholinguistic Research* (Special issue on sentence processing) 18(1). [EZ]
- Swinney, D. & Osterhout, L. (1990) Inference generation during auditory language comprehension. In: *Inferences and text comprehension*, ed. A. Graesser & G. Bower. Academic Press. [EZ]
- Swinney, D., Zurif, E. B. & Nicol, J. (1989) The effects of focal brain damage on sentence processing: An examination of the neurological organization of a mental module. *Journal of Cognitive Neuroscience* 1:25–37. [EZ]
- Tegner, R. & Levander, M. (1991) Through a looking glass. A new technique to demonstrate directional hypokinesia in unilateral neglect. *Brain* 114:1943–51. [CU]
- Teller, D. Y. (1984) Linking propositions. *Vision Research* 9:1481–90. [RS]
- Tippett, L. J. & Farah, M. J. (1992) A model of naming in Alzheimer's disease. *Bulletin of the Psychonomic Society* 30:444. [RAM]
- (in press) A computational model of naming in Alzheimer's disease: Semantic, visual, and lexical factors. *Neuropsychology*. [aMJF]
- Tulving, E. (1972) Episodic and semantic memory. In: *Organization of memory*, ed. E. Tulving & W. Donaldson. Academic Press. [aMJF]
- (1983) *Elements of episodic memory*. Oxford University Press. [aMJF]
- Umlilt, C. (1988) Orienting of attention. In: *Handbook of neuropsychology*, vol. 1, ed. F. Boller & J. Grafman. Elsevier. [CU]
- Verfaellie, M., Rapcsak, S. Z. & Heilman, K. M. (1990) Impaired shifting of attention in Balint's syndrome. *Brain and Cognition* 12:195–204. [aMJF]
- von Klein, B. E. (1977) Inferring functional localization from neurological evidence. In: *Explorations in the biology of language*, ed. E. Walker. Bradford Books/MIT Press. [aMJF]
- Warrington, E. K. (1975) The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology* 27:635–57. [BB]
- (1981) Neuropsychological studies of verbal semantic systems. *Philosophical Transactions of the Royal Society, Series B* 295:411–23. [BB]
- (1985) Agnosia: The impairment of object recognition. In: *Handbook of clinical neurology*, ed. P. J. Vinken, G. W. Bruyn & H. L. Klawans. Elsevier. [aMJF]
- Warrington, E. K. & McCarthy, R. (1983) Category specific access dysphasia. *Brain* 106:859–78. [aMJF, EZ]
- (1987) Categories of knowledge: Further fractionation and an attempted integration. *Brain* 110:1273–96. [arMJF, BB, EZ]
- Warrington, E. K. & Shallice, T. (1969) The selective impairment of auditory verbal short-term memory. *Brain* 92:885–96. [AWY, EZ]
- (1984) Category specific semantic impairments. *Brain* 107:829–54. [aMJF]
- Wasserman, G. S. (1992) Isomorphism, task dependence, and the multiple meaning theory of neural coding. *Biological Signals* 1:117–42. [GSW]
- Weiskrantz, L. (1968) Some traps and pontifications. In: *Analysis of behavioural change*, ed. L. Weiskrantz. Harper & Row. [RAM]
- Yantis, S. & Jonides, J. (1990) Abrupt visual onsets and selective attention: Voluntary vs. automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance* 16:121–34. [CU]
- Young, A. W., de Haan, E. H., Newcombe, F. & Hay, D. C. (1990) Facial neglect. *Neuropsychologia* 28:391–415. [CU]
- Young, A. W., Newcombe, F., Hellawell, D. & de Haan, E. H. F. (1989) Implicit access to semantic information. *Brain and Cognition* 11:186–209. [RvH]
- Young, M. P. & Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex. *Science* 256:1327–31. [rMJF]
- Zurif, E. B. (1980) Language mechanisms: A neuropsycholinguistic perspective. *American Scientist* 68:305–34. [aMJF]
- Zurif, E. B. & Swinney, D. (in press) The neuropsychology of language. In: *Handbook of psycholinguistics*, ed. M. Gernsbacher. Academic Press. [EZ]
- Zurif, E. B., Swinney, D., Prather, P., Solomon, J. & Bushell, C. (in press) An on-line analysis of syntactic processing in Broca's and Wernicke's aphasia. *Brain and Language*. [EZ]