BEHAVIORAL AND BRAIN SCIENCES (1997) **20,** 515–536
Printed in the United States of America

Continuing Commentary

Commentary on **Martha J. Farah (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. BBS 17:43–104.**

Abstract of the original article: 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, disengagment 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.

The fragility of the locality assumption: Comparative evidence

Philip J. Benson
University Laboratory of Physiology, Oxford, OX1 3PT, United Kingdom.
philip.benson.@physiol.ox.ac.uk; www.physiol.ox.ac.uk/~pjb

Abstract: The locality assumption (LA) seems rather awkward, especially when one considers centres of neuronal specialisation as defined by observed CNS activity. It is clear from electrophysiology that extra-striate functional compartmentalisation (modularity) is rather less well-defined than first thought; neuropsychological assessment attaching significance to varieties of preserved behaviour also reveals that some basic flaws must be inherent in current reasoning supporting LA.

If De Renzi's (1986, case 4) patient really did exhibit a pure form of prosopagnosia then Farah's (1994) hypotheses regarding the failure (inadequacies) of the locality assumption (LA) would probably have encountered considerable difficulties. It is fortunate, then, that in the face of more thorough investigations we have been able to literally dissect the nature of representation (memory) and patterns of recall (recognition). This has, however, highlighted a number of problems.

It is clear that the probability of demonstrating clearcut forms of agnosia is very low indeed. A number of reports demonstrate why this should be the case. Within the domain of neuropsychology, there is evidence demonstrating preserved post-morbid perception of other homogeneous, learned, natural objects (e.g., flowers, Campbell & de Haan 1994; cows, Bruyer & Velge 1981), yet the issue of "prosopagnosic or prosopamnesic behaviour?" remains unclear, complicated by the need to discriminate between stages of processing and levels of access to particular categories of memories. For that matter, it appears that lower species are adept at learning the appearance of novel synthesised objects (Logothetis et al. 1994), and there is a complementary wealth of psychophysical evidence demonstrating our own ability to learn, appreciate, and interact with properties of novel forms.

Important single-cell electrophysiology studies (Fuster & Jervey 1982; Tanaka et al. (1991) provide arguments against compartmentalisation of higher-order cortical function and have demonstrated the inherent difficulty in dissociating face- from other-object processing within the ventral processing stream (the temporal lobe; two distinct routes contributing to the visual processing of objects exist – for a recent summary engaged in clinical assessment see Goodale et al. 1994 and Carey & Milner commentary 1994). The first point on which I wish to concur is that it is unlikely that a recognition deficit for one class of familiar object could exist without some disruption to the visual processing of another. If this were not the case, then one would be required to conclude that separate expert systems exist and evolved in situ to deal with new classes of learned visual phenomena, a requirement which seems ludicrous given biophysical limitations in the confined space of the visual system.

It is not necessary to raise the issue of nature versus nurture regarding the apparent specificity and preferences (e.g., facial simulacra) of the higher visual system, although it is most certainly interesting to speculate why face processing might be so frequently and so markedly disrupted yet access to other, clearly feature salience-related information, may be spared. Is the reason we so rarely observe such relatively impoverished disruption of other high-level processing reflected simply by the fact that during our lifetime we spend far less time specialising in the visual appearance of other objects? Instead, surely one should pose the question whether it is reasonable to assume that one highly evolved and plastic architecture exists which can accommodate all manner of learned and even arbitrary experimental stimuli. In this respect, Caramazza et al. (1990) are quite right to argue that neuropsychological deficits/dissociations tell us nothing, or at least very little, about the nature of internal representation. Could Farah's account of memory and attentional mechanisms sympathetically embody the conspicuous complexity of inferotemporal neural processing without refuting clinical data, and if not, which aspects of her arguments would be weakened by accommodating this viewpoint?

Preserved perceptual or functional knowledge in apparently category-specific deficits is indubitably a matter for concern (Humphreys & Riddoch 1994) as it does not fit well within simulations employing fully distributed representations (cf. Burton et al. 1990, IAC versus PDP; Burton & Bruce 1994). IAC and PDP models may indeed yield the same results under particular

 0140-525X/XX $9.00+.10

operating conditions; however, independence between biological structures dealing with unique categories has to be guaranteed (as exemplified by neuropsychological evidence). An appropriate front-end to such a system may well involve rapid (early) processing of distinctiveness information (Benson & Perrett 1991; 1994; consider also temporal processing limitations due to connectivity constraints, Cowey 1985; Perrett et al. 1992). In addition, cognitive priming studies strongly suggest that distinct stages are involved in face recognition, and as such clear channels of processing are involved; their disruption produces pronounced perceptual deficits (e.g., face recognition and matching, and expression analysis, Young et al. 1993). My second question hence concerns whether norm-based distinctiveness processing is a shared (centralised) cortical function or one that is embedded within structures dedicated to type-specific object memories? The problem is that if such a "module" were disrupted, one would have to infer that inter-exemplar discrimination will suffer across all categories of objects. On the other hand, and on a more positive note, might we otherwise be in a position to examine the utility of preserved visual distinctiveness processing in agnosics?

Results from lesion and ablation studies in monkeys indicate that there is still a great deal to be understood about localisation of cortical function. For example, removal of the so-called colour processing centre, V4, does not render the subject achromatopsic; rather, colour constancy thresholds are disrupted (Walsh et al. 1993). Ablation of the "face processing area" and beyond in monkey (upper and lower banks and floor of the superior temporal sulcus, STS) markedly affects discrimination of eye-gaze, yet it does not produce a deficit in face recognition such as that observed in agnosic patients (Heywood & Cowey 1992). As Gross (1992) points out, the mere knowledge and localisation of the most complex face processing cells thus far reported does not necessarily indicate that these cells are responsible for face recognition per se. If such experiments are of use in explaining overt dysfunction then they provide rather compelling arguments against the LA and find favour with Farah's suggestion that the assumption is, indeed, "probably not correct in general."

ACKNOWLEDGMENT

Preparation of this manuscript was supported by grants from the United Kingdom Medical Research Council and the Oxford McDonnell-Pew and MRC Centres for Cognitive Neurosciences.

Locality, modularity, and computational neural networks

Horst Bischof

Department for Pattern Recognition and Image Processing, Technical University Vienna, Treiltlstr, 3/1832, Austria. **bis@prip.tuwien.ac.at; www.prip.tuwien,ac.at/**

Abstract: There is a distinction between locality and modularity. These two terms have often been used interchangeably in the target article and commentary. Using this distinction we argue in favor of a "weak" modularity. In addition we also argue that both PDP-type networks and box-and-arrow models have their own strengths and pitfalls.

In the target article and in the Author's Response (Farah 1994t, 1994r), Farah tries to demonstrate how the locality assumption might lead us in the wrong direction concerning inferences about brain organizations. Her examples using PDP models show nicely that certain phenomena are explainable without dedicated modules for specific functions. Farah has done a great job in showing that the locality assumption is a working hypothesis and may therefore be questioned. As Farah correctly notes, she introduces another set of assumptions that might be right or wrong and cannot be proved from the simulation results she has presented.

In this commentary I would like to point out several things that have been missed or misinterpreted in the whole treatment. The first thing I would like to note is that there is a difference between locality and modularity. Then I present several computational arguments in favor of a "weak" modularity assumption for PDP-type models. I will also demonstrate that this does not necessarily imply information encapsulation, which is not biologically plausible. Finally, I present some ideas on the relationships between "box-and-arrow" models and PDP-type networks, showing that each kind of description has its own strengths and pitfalls.

1. Locality. The terms modularity and locality are often used interchangeably in the treatment. Though these terms are related, they imply different things. Modularity is an abstract attribute of a system (e.g., in this particular case, an information processing device); it tells us how information is processed therein. In the case of brain organization, modularity tells us something about the connectivity; that is, there is strong interconnectivity within modules and only weak connectivity between modules.

Locality, on the other hand, tells us how information processing is physically done. In particular, locality implies that nearby locations compute similar functions (or contribute in computing the same function). Modularity is therefore solely concerned with a topological structure, whereas locality is also concerned with geometrical structure. For brains, there are some strong arguments in favor of locality (not necessarily modularity), for example, most connections are of short range and connect to nearby neurons. The minimum cost principle (see Tsotsos 1990) implies that in order to minimize connection length and maximize transmission time similar functions should be computed by neurons adjacent to each other. Following this line of argument we can say that if the cognitive architecture is modular, then it has to be local. Of course the reverse is not necessarily true. Farah's real concern is not locality (in the sense defined above), but modularity of the cognitive architecture in the sense of Fodor (1983; see also multiple book review of Fodor's *The Modularity of Mind, BBS* 8(1)1985).

2. Modularity. Let us now look at modularity from a perspective of PDP models, or more exactly, of computational neural networks (CNN; for a definition see Bezdek 1992). There has recently been a shift in the area of computational neural networks from unstructured network topologies (e.g., fully connected, 3-layer, feed-forward networks) to more structured (i.e., modular/hierarchical) ones. The main reason for this comes from computational arguments such as: faster training, better scaling behavior, improved generalization ability, and robustness and fault tolerance. (For a discussion of these features, see, for example, Bischof 1993; Jacobs et al. 1990; Jordan & Jacobs 1994.) Many of these improved capabilities of modular/hierarchical networks can be attributed to a reduction in crosstalk (either spatial or temporal; Jacobs et al. 1990), that is, conflicting information in the network. Modularity is one way to overcome these crosstalk problems. This does not imply modularity in the sense of Fodor (1983; e.g., information encapsulation). Indeed, Farah does not argue against modular networks at all (all her models have a modular structure); her main concern is information encapsulation.

What does the hypothesis of information encapsulation imply for brains and computational neural networks? Its key component is that the product of computation within one module only becomes available when the computation is finished. Imagine a neuron (or a set of neurons) that communicates with other modules. To realize the previous requirement this neuron cannot fire during the processing of the module; therefore it cannot participate in the computation of the module. Further, some additional neural machinery (e.g., inhibitory interneurons) is necessary to hinder this neuron from firing. Only when the computation inside the module is finished is the neuron allowed to fire. From this argument one can see that information encapsulation is an unrealistic assumption for brains because resources (neurons) are wasted and the minimum cost principle is violated.

3. Vocabulary of description. Many commentators have argued that PDP models are not the right way to describe cognitive phenomena. In my view, the question is not whether the descrip-

tion mechanism is right or wrong but whether it is adequate. In this respect both PDP models and box-and-arrow models have their own strengths and weaknesses.

PDP models offer an alternative, fine-grained way to describe cognitive phenomena and algorithms. The vocabulary used is that of units, connections, weights, activation functions, and so on. Such a description offers many degrees of freedom that naturally lead to learning algorithms, internal representations, and so forth. In Bischof (1993) we have presented various examples of how a description of "classical" algorithms in terms of neural networks might lead to a generalization and new insights about these algorithms. Of course, with such a fine-grained description it is hard to describe large systems. There is always a tendency to use only a few large modules and to rely heavily on learning.

On the other hand, box-and-arrow models provide a good overview and it is rather easy to describe large systems. To give a fine-grained description of a system with box-and-arrow models, however, one must use many boxes (modules). And this is what Farah points out in her target article; it is not always necessary to propose individual modules for every observed phenomenon. Good cognitive modeling should include box-and-arrow models on both a coarse scale and a fine-grained description (e.g., PDP models).

ACKNOWLEDGMENTS

This work was supported by the Austrian National Fonds zur Förderung der wissenschaftlichen Forschung under grant S7002MAT.

Neuropsychological inference using a microphrenological approach does not need a locality assumption

Wim E. Crusio

Génétique, Neurogénétique et Comportement, CNRS UPR 9074, University of Orléans, 45071 Orléans Cedex 02, France. **crusio@citi2.fr**

Abstract: Although Farah makes a convincing case against the tenability of the locality assumption, she does not propose alternative research strategies that do not rest on this assumption. It is proposed here that we may profitably exploit individual differences in neuroanatomy and behavior. In combination with the use of adequate genetic methods, this approach – termed microphrenology – does not need a locality assumption.

Farah's (1994) target article provides an admirable overview of the problems connected with the use of the locality assumption, one that more or less equates the function of a lesioned structure with the defects exhibited by the damaged brain and is almost always invoked to interpret the results of lesion studies. In an elegant way, Farah provides evidence that this reasoning may lead to false conclusions. Besides being convincing, her treatment is also constructive, in that she provides alternative hypotheses that may explain the data.

A remarkable feature of the target article (plus most of the commentaries) and, indeed, of much of neuropsychology, is the comparatively sparse use of information from other branches of neuroscience. This is all the more striking because I think that neuroanatomical and neurophysiological data provide massive support for Farah's thesis. For example, the mammalian hippocampus is a distinct structure anatomically and might easily give rise to a modular interpretation of the brain. Yet anatomical and physiological evidence clearly indicate the presence of reciprocal connections with, for example, the entorhinal cortex. These regions accordingly interact (Jones 1993), the physiological state of the one modulating that of the other.[1] These data make it impossible to assume that the functioning of the hippocampus would remain unchanged after entorhinal damage or vice versa.

Although Farah shows considerable creativity in proposing alternative explanations for a number of lesion-induced defects, no attempt is made to devise research strategies that would not rest on the locality assumption. In the field of neurobehavioral genetics such an approach already appears to exist: using genetic methods to exploit naturally occurring individual differences as a tool for understanding brain function. No brain is like another[2] and every individual behaves differently. The assumption that there is a link between the variability of the brain and individual talents and propensities appears quite plausible. This approach differs from the usual one in neuropsychology in two important respects. First, no subjects are studied that (by accident or by design) have damaged brains. Rather, all subjects will fall within the range of normal, nonpathological variation. Second, instead of comparing a damaged group with normal controls, we study a whole range of subjects and try to correlate variation at the behavioral level with that at the neuronal level. This strategy is reminiscent of the phrenological approach propagated by Franz Josef Gall (1743–1826); Lipp has coined the name "microphrenology" for it (Lipp et al. 1989). It appears that, as long as variation in one neuronal structure is independent of that in another, there will be no need for a locality assumption to interpret results of experiments carried out along these lines (contrary to Gall's own strong support for locality; cf. van Gelder 1994). In combination with methods from the field of behavior genetics (Crusio 1992), this strategy yields a very powerful approach. For example, to "magnify" individual differences, we may study animals from different inbred strains and look for correlations between the means obtained for different variables. (See Crusio et al. 1993 and references therein for some illustrative examples.) Alternatively, genetic correlations may be used to help clarify brain-behavior relationships (Crusio 1993).[3]

With the advent of noninvasive brain imaging methods such as MRI (magnetic resonance imaging) and PET (positron emission tomography), this approach is becoming increasingly feasible for use with human beings and some interesting results are already being obtained (e.g., Squire et al. 1992; see also Posner's 1994 commentary). Some of these techniques might be fruitfully applied to specific problems mentioned by Farah. For example, if we subjected a number of healthy volunteers to some tests involving their knowledge of living and nonliving things and simultaneously assessed their brain activity with PET, we would expect to see a selective activation of the temporal lobe. By appropriately manipulating test items, we might subsequently detect whether such changes correlate with the living/nonliving dichotomy proposed by Warrington and Shallice (1984) or with the visual/functional dichotomy proposed by Farah. In Warrington and Shallice's model, using knowledge of living things and nonliving things would activate different regions. Farah's model would predict that using visual information as opposed to other functional information preferentially activates different regions. If the locality assumption were correct, these differential activations would be exclusive, that is, only one region would be activated according to the property of the information needed. If the locality assumption were false, the activation would be expected in more than one region at a time but, according to the type of semantic memory implied, one region would be more strongly activated than the others.

Kosslyn and Intriligator (1992) have warned us of the perils of "sitting on a one-legged stool" and have advised neuropsychologists to use a three-legged one: like Farah, they advocate using behavioral data, computational modeling, and neural constraints to formulate and test theories. I suggest that a four-legged stool would be even more solid: let us add the study of individual differences in brain and behavior to the neuropsychological chair.

ACKNOWLEDGMENTS

This study was supported by the Centre National de la Recherche Scientifique (UPR 9074), Ministry for Research and Technology, Région Centre, and Préfecture de la Région Centre. UPR 9074 is affiliated with INSERM and the University of Orléans.

NOTES

1. Another example, the functional overlap between the hippocampus and the superior colliculus, has been discussed in detail by Foreman and Stevens (1987). By "functional overlap," these authors meant that "effective functioning of structure A is dependent on information handling in structure B, or vice versa, that the involvement of both structures is necessary for efficient performance of a particular behaviour" (p. 102).

2. This heritable variation of the brain is another aspect that neuropsychologists (and many neuroscientists as well) tend to ignore, most likely to their own peril. For example, Donovick et al. (1981) and Fanelli et al. (1983) reported widely divergent behavioral effects of septal lesions in mice, depending on which particular inbred strain was being used.

3. The defects shown by neurological mutants can be regarded as the neurobehavioral-genetic parallel of lesions. Such animals may solve another practical problem inherent in many lesion studies: almost no lesion is limited to only a single brain structure, not even in controlled animal studies. Animals with an intact dorsal commissure of the fornix but without a corpus callosum are almost impossible to obtain by means of a surgical intervention. However, appropriate crosses between the inbred mouse strains BALB/cWahl and 129/ReJ, for example, may produce just such animals (Wahlsten & Schalomon 1994). Still, neurological mutations are seldom limited in their effects to one brain structure only, so this approach constitutes more a (welcome) addition rather than a real alternative to the lesion technique.

The "locality assumption": Lessons from history and neuroscience?

Jonathan K. Foster

Department of Psychology, University of Manchester, Manchester, M13 9PL, United Kingdom. **foster@psy.man.ac.uk; www.psy.man.ac.uk**

Abstract: This commentary seeks to place Farah's (1994) arguments in the historical context of ideas about mind-brain relationships. It further seeks to draw a conceptual parallel between the issues considered by Farah in her target article and questions which have concerned neuroscientists since the nineteenth century regarding the functional organization of the brain. Specific reference is made to the relationship between use of the concept of "locality" in cognitive neuropsychology and use of the concept of "localization" in neuroscience.

It is a truism that in psychology what goes around historically tends to come around. Witness the rehabilitation of connectionism after many decades in the intellectual wilderness, almost a century after Edward Thorndike first coined the term. This revolutionary trend is especially apparent when one considers the question of localization of cognitive function. Almost as soon as the phrenologists were producing their highly detailed, local functional maps of the cortex, their equipotentialist opponents were arguing that such a precise mapping of functions was not feasible. The debate has continued ever since.

In her critique, Farah (1994) goes some way in acknowledging the historical context of the debate surrounding this issue, although it seems somewhat ironic that the author quotes from David Ferrier (1886) for, at the neural implementation level, the nineteenth century electrophysiological work of Ferrier and his contemporaries is usually interpreted as supporting the localizationist rather than the distributionist position. Indeed, the subsequent antilocalizationist stance of Goltz represented a direct challenge to the position of Ferrier, Eduard Hitzig, and Gustav Fritsch.

This viewpoint was subsequently echoed in the work of Karl Lashley, whereas, more recently, other researchers have proposed that information may be represented as patterns distributed across multiple locations, using, for example, mechanisms analogous to the storage of information in holograms (Pribram 1971; 1982). In many respects, this kind of approach foreshadowed more recent developments in computational approaches to neuropsychology, such as PDP (Hinton & Anderson 1981; Rumelhart & McClelland 1986).

These considerations notwithstanding, Farah's target article makes a significant contribution to this ongoing debate, providing a timely examination of the veiled, yet pervasive, "locality assumption" of contemporary cognitive neuropsychology. With admirable clarity of exposition, the author presents alternative distributed and interactive ways of conceptualizing evidence from the cognitive neuropsychology literature in the areas of semantic memory, attention, and face recognition, making the bold claim that assumptions of modular locality, whether tacit or explicit, may, in certain cognitive domains, be fundamentally flawed. Hard-nosed localists in the neuropsychology fold should take careful note that Farah's thesis and the evidence adduced in support of her argument. It is important that, as the author delineates, it is now possible to make these questions explicit through computational models. Farah's framework promises to be fruitful. Furthermore, it provides considerable heuristic potential through the formulation of clear-cut, testable predictions.

Outstanding questions remain, however, for those who embrace the nonlocality line. For example, in the neuropsychology of memory, considerable evidence has been amassing over the past few decades that focal, specific lesions of particular brain structures (most notably, the hippocampus) can produce profound, enduring, and somewhat "local" deficits in mnemonic function (see Squire 1992 for a review). In the future, theorists and experimentalists alike will need to make vigorous attempts to reconcile conflicting distributionist and locationist tensions arising from findings in memory and other cognitive domains.

A further question prompted by Farah's article concerns the relationship between the relatively recent notion of locality and the historically older question of localization of function within the substrate of the brain: specifically, the question of the implications of the latter for the former. I take locality to refer to the ascription of circumscribed function at the level of the cognitive architecture, and localization to refer to the attribution of regional function at the level of the neural substrate. Historically, the latter has been the province of what I shall call neurological neuropsychologists, whereas the former has, more recently, become the domain of cognitive neuropsychologists. Although a comparison of the two may leave one open to the criticism of "confusing levels of explanation," one plea in defence is that the vast majority of cognitive neuropsychologists appear to subscribe, at least implicitly, to the notion of some degree of localization of cognitive function within the wetware of the brain. Fodorian modularity theory would also appear to be implicitly localizationist.

Furthermore, there is a significant conceptual parallel between the question of the validity of the locality assumption and the usefulness of the lesion method in physiological psychology. (There is insufficient space to do this debate justice here; the interested reader is referred to Gregory [1961], Webster [1973], and Weiskrantz [1968; 1974] for discussions of theoretical issues, and to Schoenfeld & Hamilton [1977] for an overview of practical problems associated with the lesion methodology.) In brief, the fundamental rationale of lesion studies in physiological psychology is that the function of an area of the brain can be inferred from the behavioural or psychological capacity or capacities that are absent from the organism's repertoire after destruction of a particular brain region. This rationale, in turn, rests on the fundamental tenets of lesion research as follows: (1) behavioural functions are represented in discrete brain structures, such as nuclei and fibre tracts and (2) lesions disrupt function by removing functional tissue in circumscribed sites in the brain. However, these tenets have been challenged in the physiological psychology and neuropsychology literature on both theoretical and practical grounds. One is reminded of Goldstein's famous injunction that what is important is not so much the brain injury *per se,* as the response of the remaining part of the system to that injury (see Heeschen, 1985, p. 209). One can also discern a similarity between Farah's thesis and interactionist theory, derived from the work of Hughlings-Jackson, who believed that higher-level functions are founded on a number of lower-level processes. According to this approach, it may be possible to localize these more basic component skills quite accurately, but they are combined to generate

higher-level cognitive processes in flexible ways. Alternatively, the critical questions at the systems level may be the size of the anatomically defined functional area across which information is distributed, and how broad a class of information may be equivalently represented in a particular region of neural tissue. There may be some form of regional equipotentiality within the cortex, but only within relatively circumscribed, well-defined regions.

A further point concerns the domain of interest in cognitive neuropsychology. Historically, neuropsychologists have tended to think about the cortex and subcortex quite separately, even though they are intimately connected at the neuroanatomical and neurochemical levels. Whereas the cerebral cortex is typically studied in humans, the subcortex has been largely studied in animals. The quotation from Ferrier (1886) at the beginning of Farah's piece refers to allocation of function within the encephalon or whole brain. However, contemporary cognitive neuropsychology, as espoused by Farah, tends to concern itself predominantly with cognitive deficits following focal damage to the cerebral cortex. Indeed, one might argue with some cogency that the emphasis on locality derives from subscribing to the notion of localization of function (whether implicit or explicit), combined with studying patients whose brain lesions are predominantly focal (real or imagined). However, one should not neglect possible subcortical involvement in higher-level cognitive processes. In addition to being involved in memory, subcortical structures have been clearly implicated in the mediation of such high-level cognitive processes as language and attention. Subcortical regions may exert a neuromodulatory effect on cognitive function via their diverse projections. This modulation can be conceptualized in terms of a distributed (nonlocal, in modular terms?) "irrigation system" innervating higher cortical regions.

Finally, one of Farah's central arguments, voiced near the beginning of her critique (sect. 1.1, para. 1), is that the notion of locality follows incontrovertibly from an adherence to the Fodorian concept of modularity. Furthermore, Farah seems to regard the locality assumption as axiomatic in contemporary cognitive neuropsychology thereby conveying the impression that she regards the concept as one "hard core" feature (Lakatos 1974) of the entire body of scientific enquiry. The notion of locality would certainly seem to predate Fodor in the theories and writings of the phrenologists and their acolytes. However, reexamination of Fodor's major written output (Fodor 1983; 1985), suggests that modular locality can be inferred only indirectly from Fodor's seminal works. For example, modular systems are directly described as being "domain specific, innately specified, hard-wired, autonomous, and not assembled" (1983, p. 37). Further, Fodor contends that modules are informationally encapsulated, that their processes are mandatory and rapid, and that their output is shallow. The word "local" is conspicuously absent, other than when, in his later Précis, Fodor (1985) describes central processes as the antithesis of modular systems, being "slow, deep, global *rather than local* [my emphasis], largely under voluntary (or, as one says "executive") control, typically distributed with diffuse neurological structures" (p. 4). Given its central importance to the debate, this point requires further classification and elaboration, if not from Farah, then from Fodor himself.

ACKNOWLEDGMENTS

I am indebted to Cynthia McDonald, Department of Philosophy, University of Manchester for sharing her thoughts on the theories of Marr, Fodor, modularity, and classical versus connectionist models of cognition.

What is the locality assumption and how is it violated?

Vinod Goel,[a] Paolo Nichelli,[b] and Jordan Grafman[c]

[a]*Department of Psychology, York University, North York, Ont., Canada M3J 1P3.* **vgoel@yorku.ca.** [b]*Clinica Neurologica, Modena, Italy.* **nichelli@imoaxl.unimo.it.** [c]*Cognitive Neuroscience Section, National Institute of Neurological Disorders & Stroke, Bethesda, MD 20892-1440.* **jgr@box-j.nih.gov**

Abstract: We respond to Farah (1994) by making some general remarks about information encapsulation and locality and asking how these are violated in her computational models. Our point is not that we disagree, but rather that Farah's treatment of the issues is not sufficiently rigorous to allow an evaluation of her claims.

Farah (1994) raises some important and timely issues in the target article. Unfortunately, her treatment of these issues lacks the precision and substance which inform her empirical work, and serves more to obscure than to clarify. We will restrict ourselves to some general remarks about information encapsulation and locality and the specific question of how they are violated in Farah's computational models. Our point is not that we disagree with her, but rather that her treatment of the issues is not sufficiently rigorous to allow an evaluation of her claims.

Farah takes as her starting point Fodor's (1983) notion of information encapsulation. She states that information encapsulation *implies* locality (or locality follows from information encapsulation; pp. 43–44) and then goes on to argue that the locality assumption is probably false. We are certainly very sympathetic with her primary motive for arguing thus: namely, it makes it too easy to reify behavioral deficiencies as part of the functional architecture. However, we do not understand how locality follows from information encapsulation. Furthermore, we can make some (limited) sense of what the claims of information encapsulation and locality amount to in standard computational systems but we are not at all sure what they amount to in connectionist networks.

There is a consensus that neuropsychology and cognitive neuropsychology are in the business of articulating the "functional architecture" of the normal mind by examining the behavioral consequences of neurological pathologies. There is also some appreciation of the fact that for early neuropsychologists, like Gall, articulating the "functional architecture" meant carving up the neurophysiological mechanism at its causal joints, while for cognitive neuroscience it means carving up the computational system at its "informational joints." Given that this is the intellectual endeavor we are engaged in, what are some of the assumptions we need to make? Here there is also consensus that we need to assume the system whose structure we are trying to induce is modular to some extent (McCarthy & Warrington 1990; Shallice 1988; Vallar 1991).[1] There are few clear statements, however, apart from Fodor's (1983; see also multiple book review of Fodor's *The modularity of mind. BBS* 18(1) 1985), about the differences in the notions of modularity required by neuropsychology and *cognitive* neuropsychology.

For pioneers like Gall, modularity was the belief that one could individuate behaviorally distinct cognitive functions and map them onto causally distinct neurophysiological subsystems without cross-classification. Both neurophysiological structures and lesions have a specific location in space. One can measure both the extent and the location of damage and its behavioral consequences. Of course, one is not interested in spatial location per se, but rather the causal joints. But given a nineteenth century conception of how the world works (e.g., no action at a distance and determinism), it was not unreasonable to associate causal contiguity with spatial contiguity and locality (though one neither implies nor is implied by the other).[2] In this world view the distinction between causal and spatial contiguity can be overlooked (on the assumption that there will be a clean mapping between the two) and the search for modules becomes the project

of mapping behavioral deficits onto localized neurophysiological structures.

Life is much more complicated for contemporary *cognitive* neuropsychology. If we take the cognitive conception of the world seriously and believe that any generalizations that do justice to human behavior will have to causally involve the semantic or information content of our mental states, we are obligated to map our behaviorally individuated functions onto computational structures and procedures individuated along what one might call "flow of information" lines. This is the level and notion of modularity captured by information encapsulation (Fodor 1983). The complicating factor is that, whereas the computational structures and processes are underwritten by the neurophysiological structures, the well known multiple realizability results regarding computational systems are widely interpreted as allowing few (if any) inferences from the structure of computational procedures to the structure of the mechanism which realizes the procedures, or vice versa (Fodor 1975; Newell 1980; Pylyshyn 1984).[3] This has several consequences. First, it means that, as good cognitive neuropsychologists, we need not be particularly interested in the causal joints in the physiological structures. Second, even if we are interested, we can get little mileage out of spatial localization on the old assumptions. To associate computational procedures with spatial location we now need to make the strong assumption that (1) the individuation of computational procedures along "informational joints" does not cross-classify the individuation of neurophysiological subsystems along their causal joints and that (2) causal contiguity corresponds to spatial contiguity in these subsystems. These additional assumptions, contrary to Farah's claim, are certainly not implied or required by information encapsulation, and there are some unanswered questions about their compatibility with, and effect on, the computational theory of mind. The situation is further complicated by the introduction of certain connectionist models where the clear distinction between hardware and software is blurred and there is still no understanding of the sense in which they are doing information processing (Cummins 1989; Goel 1991). In these cases it is very difficult to say what information encapsulation and locality might amount to, and what their relationship might be.

This state of affairs requires very careful, specific statements about the significance, meaning, and relationship of locality and information encapsulation in both classical computational models and connectionist networks. The latter is the more urgent because we have little understanding of these issues in such computational systems. Farah devotes the bulk of her efforts to presenting connectionist models of three well known dissociations and making the argument that these models provide explanations of the phenomenon as good as or better than the standard interpretations, *and that they do so by virtue of violating the locality assumption.* We briefly consider the first of these models ("the functional architecture of semantic memory"), though our comments are intended to generalize over the others.

Farah's model is an instance of an auto-associator network (McClelland & Rumelhart 1986; 1989). It contains three "pools" of nodes. Within each pool, each node is connected to every other node, resulting in n^2-n connections. There are (presumably) n connections between pools. One pool is called the "visual input," another the "verbal input," and the third is the "semantic memory." Within semantic memory, one-third of the units represent functional information while two thirds represent visual information. The relevant claim with respect to the model is the following (p. 50): "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."

We do not argue that this claim is incorrect, but rather, we simply do not know how to evaluate it. Is the assumption that the representation of functional and visual information in semantic memory of the model constitutes different modules, but a lesion to the visual semantics also effects the functional semantic, thus violating locality? There is a natural sense in which the three pools of nodes constitute modules (by virtue of intra-pool connections being much denser than inter-pool connections). However, it is not clear what notion of modules is the relevant one. Are the interconnections to be interpreted as physical, informational, or both? Furthermore, if these pools constitute modules along some dimension, what notion of modularity is Farah appealing to in claiming that the units representing functional and visual memory constitute modules? There is no variability in the density of interconnections between them. Given our understanding of the role and structure of computational explanations in cognitive science and our reading of her text, we do not believe that Farah has confronted the various complex issues implicated by her claim.

In summary, we find Farah's treatment of modularity exasperating because it is too fast, conflates a number of distinct issues, and rides roughshod over others. But we do commend her for introducing connectionist modeling into the discussion. Despite her denials of comparing apples and oranges, the difficulty of interpreting some connectionist networks and the uncertainty of the relationship between physical and computational organization introduces a new dimension into the discussion and raises interesting questions about the relationship of Gall's and Fodor's notions of modularity.

NOTES

1. We do not deal here with the question of whether modularity is actually required to draw such inferences. We are only interested in clarifying what the modularity assumption is.

2. On a twentieth century conception of the world (indeterminism, chaos theory, etc.), this is a much less secure assumption.

3. One of us (Goel 1991; 1992) has tried to argue that there are some minimal inferences one can draw from the structure of computational systems to the structure of the implementing mechanisms, but this is very much a minority position.

ERPs and the modularity of cognitive processes

Valerie Gray Hardcastle

Department of Philosophy, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0126. **valerie@vt.edu; http://mind.phil.vt.edu/**

Abstract: Farah argues that nonlocal models explain clinical data better. However, the locality assumption does not seem so implausible if different sorts of data are taken into account. In particular, priming experiments using evoked response potentials support modularity. I describe some ERP studies relevant to this issue.

Farah (1994) argues that the evidence for the locality assumption in psychology and neuroscience is actually better explained using nonlocal models. However, her arguments rely largely on only two types of data: reaction time experiments from psychology and lesion studies from neurology. I suggest that there is a third important type of evidence that should not be overlooked, because it is perhaps the best bridge between psychology and neuroscience: evoked response potentials (ERPs).[1] I can, of course, produce no conclusive evidence for resolving the modularity issue (especially in this short commentary); however, what I do hope to present here is the general flavor of the type of data that should also be included in these metatheoretical discussions. These data suggest that the locality assumption is the correct approach after all.

Much of the priming research in cognitive psychology focuses on the influence the semantic properties of one word have on the subsequent recognition of later words. It has repeatedly been shown that words are recognized faster if they have been semantically primed by an earlier presentation of a related word. One avenue of investigation in semantic priming paradigms has been to

use ERPs to visual and auditory stimuli, which give a much higher temporal resolution to various aspects of cognitive processing than do any purely behavioral measure.

In general, when the ERP waveforms for semantically related pairs of words are compared with semantically different pairs of words in the visual domain, there is a difference in a late negative component with onset at around 200 msec and peak near 400 msec after the stimulus presentation, bigger over the right hemisphere than the left, and concentrated in the centro-posterior portions of the brain (see Holcomb & Neville, 1991, for discussion). This waveform is generally referred to as the N400 wave. Study under different probability conditions for semantic relatedness suggests that the more two words are unrelated, the larger the N400 waveform will be. This type of study has been repeated in the auditory domain, with similar results, although the waveform had an earlier onset, a later peak latency, and a somewhat different scalp distribution.

Most important, ERP measurements of related priming phenomena indicate specialization with respect to other domains. For example, Barrett and Rugg (1989; Barrett et al. 1988) examined identity and semantic priming in faces. They found an early negative component similar to the N400 indexes primed versus unprimed faces. This waveform, however, an N250, is larger over the frontal and parietal regions, whereas the N400 is larger over the centro-posterior regions. Moreover, the N250 shows no hemispheric asymmetries, whereas the N400 is larger over the right hemisphere (cf Holcomb & Neville 1991). These results suggest that the N250 and the N400 are elicited by different brain systems even though they are correlated with priming by previous context.

In addition, Helen Neville has gathered data from ERP studies indicating that accessing an "implicit" memory system and accessing the memory system that apparently underwrites explicit conscious experience (Neville & Weber-Fox 1994) give rise to qualitatively different kinds of ERP wave. Neville et al. examined the ERP waves for distinct patterns of priming when the general episodic system would be activated and the subjects aware of the primes, as contrasted with trials in which the primes were masked and subjects were then unaware of them. As has been well documented, she found N400 effects correlated with the explicit semantic priming. In contrast, an enhanced negativity that occurred around 200 msec after target presentation lasted only 120 msec, and had a centro-anterior distribution that marked the masked semantic priming effects. The different timing and distribution of the masked and unmasked effects suggest that nonidentical systems are being activated in the two conditions.

Moreover, one can get the N400 priming effect for both words and pseudo-words in the masked condition. This result suggests that the early anterior priming effect may indeed index access to a more specialized structural representation of words in some sort of lexicon before a more general episodic system is activated. The earlier onset and shorter duration are compatible with an automatically activated system, and these factors plus the different scalp distributions of the two priming effects point to distinct processes within the language processing system.

Our brief foray into the ERP literature suggests that at least three "modules" are present in our brain: a conscious language processor separate from a conscious face recognizer and both of these separate from an unconsciously accessed lexicon. Though I do not take the modularity issue to be thereby settled, I do wish to suggest that more needs to be included in the debate and that the types of models Farah has developed are not adequate for all known data.

NOTES

1. ERPs are electroencephalographic recordings time-locked to a series of stimuli and then averaged across like trials. Simple EEG waves contain much noise, but if several trials of the same stimulus are averaged together, the noise drops out and a waveform distinctive for that stimulus remains. Manipulating the conditions under which stimuli are given can suggest, among other things, the location of different types of processing. See Näätänen: "The Role of Attention in Auditory Information Processing as Revealed by Event-Related Potentials and Other Brain Measures of Cognitive Function" *BBS* 13(2) 1990 and Verleger: "Event-Related Potentials and Cognition." *BBS* 11(3) 1988.

Author's Response

More interactions on the interactive brain

Martha J. Farah

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

Abstract: The central claim of my original target article was a modest one (that modularity does not always hold) but it was misinterpreted as a much stronger one (that modularity never holds). Further confusions arose from multiple valid usages of the term "modularity" and the similarity of the terms "locality" and "localization." Despite the limited nature of the claim, I maintain that it poses a stubborn problem for neuropsychology, not to be dispelled by new empirical methods or a priori reasoning.

If different components of the cognitive architecture are "informationally encapsulated," that is, interact relatively little with one another in the course of performing their functions, then the effect of damage to one component will be localized to that component. In this case the behavior of the patient after brain damage will be easy for neuropsychologists to interpret, as it results from the absent or impaired functioning of the damaged component against a background of a normally functioning residual system. Informational encapsulation was Fodor's (1983) prime criterion for modularity, so the commentators and I frequently refer to such systems as modular. In contrast, if the architecture is highly interactive, with much give-and-take among components as they carry out their computations, then the effect of damage to one component will not be localized to that component but will encompass all other components that normally depend on input from it. In this case patient behavior will be more difficult to interpret, as it results from a combination of the absent or impaired functioning of the damaged component and the altered functioning of the other components of the system. In my target article (Farah 1994t), I dubbed the assumption that the effects of damage are local to the damaged component "the locality assumption," and questioned whether it was correct.

R1. An extremely nonextreme claim. It seems that many readers understood me to be saying that brain architecture is uniformly interactive, and that the effects of local brain damage are never functionally local. In fact, my target article said something much less extreme (see sect. 3.3.3). Let me reiterate the analogy I offered in response to the first round of commentary: from the fact that some cats are black, it would be wrong to assume that all cats are black. Just because some neural systems may operate in a relatively modular fashion, it would be wrong to assume that all do. In demonstrating that interactive architectures do a better job of explaining some neuropsychological data than

modular architectures I was providing examples analogous to some nonblack cats. But from these examples it does not follow that no neural systems are modular, or that no cats are black. In both my target article and my response (Farah 1994r) to the first round of commentaries, I explicitly endorsed the idea that modularity may hold in some cases.

The tendency to extremize positions must be a very robust cognitive phenomenon, because three of the new commentaries bring up apparent exceptions to interactivity as exceptions to my claim. **Benson** discusses the possibility that prosopagnosia is a face-specific disorder, with no necessary accompanying impairment in object recognition, as a potential violation of my critique of the locality assumption. **Foster** points to the selectivity of new learning impairments after hippocampal damage as inconsistent with my claim. And **Hardcastle** devotes her entire commentary to reviewing a set of event-related potential (ERP) findings that she interprets as evidence against interactivity. I am therefore grateful for the opportunity to state clearly again my more modest claim: there are cognitive functions that appear to be implemented in a highly interactive fashion; we therefore cannot assume locality or (for present purposes, equivalently) modularity.

R2. The search for a more decisive methodology. One goal of my target article was to show that the interpretation of neuropsychological dissociations is not straightforward once we begin to think about interactive architectures. In a modular system, an impairment in knowledge of living things implies damage to a "living things module." This, in turn, implies the existence of such a module, which is of interest to cognitive neuroscientists concerned with the structure and function of the normal brain. But as I argued in the target article, once the door has been opened to interactivity, we see that alternative interpretations of such data are possible. In the case of impaired knowledge of living things, an underlying loss of visual knowledge paired with certain additional assumptions can also account for the available data (see sect. 2.1.3 of the target article).

Upon discovering that neuropsychological dissociations are so ambiguous, most of us immediately start a mental search through all the methodologies we know, in the hope that we will find one that is less ambiguous. Although I agree with **Crusio** that the more methods we use, the greater our inferential power, I do not believe that there is any method or even combination of methods that will give us the generic and straightforward inferences that are possible from dissociations under the locality assumption (i.e., observe a selective impairment in X, infer an X module). The advantage of multiple methods is their complementary strengths and weaknesses, and in specific cases these can disambiguate the data and deliver a more specific inference. For example, **Crusio** urges the use of neuroimaging methods. We could indeed support or disconfirm the hypothesis that this apparently category-specific impairment is really a modality-specific impairment of visual knowledge by finding that normal subjects activate visual association cortex when performing semantic judgements about living things. The relevance of neuroimaging data here hinges on our prior knowledge of the localization of visual processing areas in the brain and the fact that one of the alternative hypotheses concerned vision. It is not clear how neuroimaging can disambiguate dissociations in the general case. The same is true for the study of individual differences: if knowledge of living things were a dimension of individual variation, we would not know whether the underlying cognitive difference was visual or category-specific.

Hardcastle reviews a number of ERP studies of language, perception, and memory. The appearance of certain ERP components in some task contexts and not others can be viewed as a type of dissociation. Hardcastle interprets these dissociations as support for modularity. In my target article I argued that dissociations are ambiguous unless one assumes modularity. It seems to me that Hardcastle is assuming modularity in interpreting the ERP data, rather than using the ERP data to infer modularity. If there is something about ERP dissociations that makes them less ambiguous than dissociations following brain damage, Hardcastle has not said what it is.

R3. Some issues are definitional. I suspect that some of the disagreements raised in these commentaries stem from different senses of words such as "modularity" and "locality." For example, **Hardcastle** refers to ERP evidence of specialization for word and face processing in the course of defending modularity. As I tried to make clear in my target article and response, the relevant sense of modularity for the issue under discussion is informational encapsulation, not specialization. Alas, with other valid usages to be found in the literature, it is not surprising that such confusions occur.

The term "locality" has also been subject to misunderstanding because of its similarity to "localization," which refers to the anatomical segregation of neurons with the same function in the same location. **Foster** contributes a helpful discussion of the relation between the two concepts. **Bischof,** on the other hand, seems to have interpreted "locality" as "localization"; he makes several excellent points about the relation between localization and modularity, but the same points do not hold for the locality assumption and modularity. **Goel et al.** seem to be warning us that localization (an anatomical concept that they illustrate by reference to Gall's phrenology) and locality (an informational concept, in their terms) are different. To reiterate: this is my view too, and is consistent with the original claims of the target article.

Goel et al. also express uncertainty about the applicability of information-processing concepts such as informational encapsulation and locality to connectionist networks. But they do not go beyond asserting that "it is difficult to say what [such concepts] might amount to" in connectionist systems, and saying of the one connectionist example that they discuss "we do not claim that it is incorrect, but rather, we simply do not know how to evaluate it." Perhaps with more space to lay out their arguments they could justify their concerns. Conceptual analysis is a valuable contribution that the philosophically inclined can make to empirical science. However, simply asserting that this is unclear and that is uninterpretable is conceptual stonewalling, not conceptual analysis.

Perhaps some etymology will help in distinguishing "locality assumption" and "localization." In my first draft of the target article, I criticized the "transparency assumption," a term coined by Caramazza in the 1980s. My understanding of this term was based on statements such as "This assumption essentially says that the cognitive system of a brain-damaged patient is fundamentally the same as that of a

normal subject except for a 'local' modification" (Caramazza 1986), and that the assumption is violated if "the remaining, unimpaired processes work differently when one component is not functioning normally" (Caramazza 1984). Note that these definitions are equivalent to the locality assumption as defined by me (e.g., in the first paragraph of the present response). In later writings, however, Caramazza revised or clarified the definition of transparency, specifying only that the behavior of the system should be understandable after damage: "hypothesized modifications of the normal processing system must be tractable within the proposed theoretical frameworks" (Caramazza 1992). Thus, I was left without a term for the assumption I wished to discuss, and I was forced to invent a new piece of terminology. "Locality assumption" seemed mnemonic, as it referred to the assumption that the effects of brain damage are local to the damaged component.

R4. And some are empirical. Two of the commentators agree with my conclusion that the locality assumption is wrong, and in fact seem to hold the even stronger belief that there are no modular systems in the human brain. Part of their rationale for this rather extreme position comes from *a priori* computational considerations. I would like to suggest that such considerations be used with caution. We do not know enough about the computational pressures on and resources of the brain to be confident of such *a priori* arguments.

An example of modularity discussed by **Benson** is the possibility that face and object recognition proceed independently of one another. He calls the possibility of such specialized and independent subsystems "ludicrous, given biophysical limitations." Yet as he points out, the phenomenon of prosopagnosia is certainly suggestive of such modularity, and recent experimental results with both selectively impaired and selectively preserved face recognition imply that we should consider this hypothesis seriously (see Farah 1996). There is as yet no alternative hypothesis that can explain both prosopagnosia and object agnosia sparing faces in terms that do not include independent components of visual recognition.

Also relying on *a priori* considerations, **Bischof** argues that informational encapsulation could not be used in the brain because it is wasteful of neural resources. His argument is valid only insofar as there are no more efficient ways of implementing informational encapsulation than he has thought of, and only insofar as the advantages of informational encapsulation do not outweigh their costs.

Ironically, it was *a priori* computational reasoning that Fodor used to argue that perception is informationally encapsulated:

> To the extent that input systems are informationally encapsulated, of all the information that might in principle bear upon a problem of perceptual analysis, only a portion (perhaps only a quite small and stereotyped portion) is actually admitted for consideration. . . . If there is a body of information that must be deployed in perceptual identifications, then we would prefer not to have to recover that information from a large memory, assuming that speed of access varies inversely with the amount of information that the memory contains (Fodor 1983, p. 70).

Horgan (1996) has claimed that we have reached "the end of science," with only the details and applications of science to be worked out. Whatever the status of this claim for the physical sciences, it surely does not apply to the current state of computational neuroscience. Our ignorance of the neural mechanisms of thought is deep; we are in no position to answer questions on the basis of first principles. Whether the computational architecture of the brain is modular or interactive can only be discovered by empirical research.

References

Barrett, S. E. & Rugg, M. D. (1989) Event-related potentials and the semantic matching of faces. *Electroencephalography and Clinical Neurophysiology* 60:343–55. [VGH]

Barrett, S. E., Rugg, M. D. & Perrett, D. I. (1988) Event-related potentials and the matching of familiar and un-familiar faces. *Neuropsychologia* 26:105–17. [VGH]

Benson, P. J. & Perrett, D. I. (1991) Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology* 3:105–35. [PJB]

(1994) Visual processing of facial distinctiveness. *Perception* 23:75–93. [PJB]

Bezdek, J. (1992) On the relationship between neural networks, pattern recognition, and intelligence. *International Journal of Approximate Reasoning* 6:85–107. [HB]

Bischof, H. (1995) *Pyramidal neural networks.* Erlbaum. [HB]

Bruyer, R. & Velge, V. (1981) Unilateral cerebral lesion and disturbance of face perception: Specificity of the deficit. *Acta Neurologica Belgica* 81 (6):321–32. [PJB]

Burton, A. M. & Bruce, V. (1994) Local representations without the locality asumptions. *Behavioral and Brain Sciences* 17(1):62–63. [PJB]

Burton, A. M., Bruce, V. & Johnston, R. A. (1990) Understanding face recognition with an interactive activation model. *British Journal of Psychology* 81:361–80. [PJB]

Campbell, R. & de Haan, E. H. F. (1994) Developmental prosopagnosia: A functional analysis and implications for remediation. In: *Cognitive neuropsychology and cognitive rehabilitation,* ed. M. J. Riddoch & G. W. Humphreys. Erlbaum. [PJB]

Caramazza, A. (1984) The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and Language* 21:9–20. [rMJF]

(1986) On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single case studies. *Brain and Cognition* 5:41–66. [rMJF]

(1992) Is cognitive neuropsychology possible? *Journal of Cognitive Neuroscience* 4:80–94. [rMJF]

Caramazza, A., Hillis, A. E., Rapp, B. C. & Romani, C. (1990) The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology* 7:161–90. [PJB]

Carey, D. P. & Milner, A. D. (1994) Casting one's net too widely? *Behavioral and Brain Sciences* 17(1):65–66. [PJB]

Cowey, A. (1985) Aspects of cortical organisation related to selective attention and selective impairments of visual perception. In: *Attention and performance XI,* ed. M. I. Posner & O. S. M. Marin. Erlbaum. [PJB]

Crusio, W. E. (1992) Quantitative genetics. In: *Techniques for the genetic analysis of brain and behavior: Focus on the mouse. Techniques in the behavioral and neural sciences, volume 8,* ed. D. Goldowitz, D. Wahlsten & R. Wimer. Elsevier. [WEC]

(1993) Bi- and multivariate analyses of diallel crosses: A tool for the genetic dissection of neurobehavioral phenotypes. *Behavior Genetics* 23:59–67. [WEC]

Crusio, W. E., Schwegler, H. & Brust, I. (1993) Covariations between hippocampal mossy fibres and working and reference memory in spatial and non-spatial radial maze tasks in mice. *European Journal of Neuroscience* 5:1413–20. [WEC]

Cummins, R. (1989) *Meaning and mental representation.* The MIT Press. [VG]

De Renzi, E. (1986) Current issues on prosopagnosia. In: *Aspects of face processing,* ed. H. D. Ellis, M. A. Jeeves, F. Newcombe & A. W. Young. Martinus Nijhoff. [PJB]

Donovick, P. J., Burright, R. G., Fanelli, R. J. & Engellenner, W. J. (1981) Septal lesions and avoidance behavior: Genetic, neurochemical and behavioral considerations. *Physiology and Behavior* 26:495–507. [WEC]

Fanelli, R. J., Burright, R. G. & Donovick, P. J. (1983) A multivariate approach to the analysis of genetic and septal lesion effects on maze performance in mice. *Behavioral Neuroscience* 97:354–69. [WEC]

Farah, M. (1994) Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences* 17(1):43–61. [HB, WEC, rMJF]

(1996) Is face recognition "special"? Evidence from neuropsychology. *Behavioral Brain Research* 76:181–90. [rMJF]
Fodor, J. A. (1975) *The language of thought*. Harvard University Press. [VG}
(1983) *The modularity of mind: An essay on faculty psychology.* MIT Press. [HB, JKF, VG, rMJF]
(1985) Précis of *The modularity of mind*. *Behavioral and Brain Sciences* 8:1–42. [JKF]
Foreman, N. & Stevens, R. (1987) Relationships between the superior colliculus and hippocampus: Neural and behavioral considerations. *Behavioral and Brain Sciences* 17:101–52. [WEC]
Fuster, J. M. & Jervey, J. P. (1982) Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *Journal of Neuroscience* 2:61–75. [PJB]
Goel, V. (1991) Notationality and information processing mind. *Minds and Machines* 1(2):129–65. [VG]
(1992) Are computational explanations vacuous? In: *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum. [VG]
Goodale, M. A., Jakobson, L. S., Milner, A. D., Perrett, D. I., Benson, P. J. & Hietanen, J. K. (1994) The nature and limits of orientation and pattern processing supporting visuomotor control in a visual form agnosic. *Journal of Cognitive Neuroscience* 6 (1):46–56. [PJB]
Gregory, R. L. (1961) The brain as an engineering problem. In: *Current problems in animal behaviour,* ed. W. Thorpe & O. L. Zangwill. Cambridge University Press. [JKF]
Gross, C. G. (1992) Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London* B335(1273):3–10. [PJB]
Heeschen, C. (1985) Agrammatism versus paragrammatism: A fictitious opposition. In: *Agrammatism,* ed. M. -L. Kean. Academic Press. [JKF]
Heywood, C. & Cowey, A. (1992) The role of the "face-cell area" in the discrimination and recognition of faces by monkeys. *Philosophical Transactions of the Royal Society of London* B335(1273):31–38. [PJB]
Hinton, G. E. & Anderson, J. A. (1981) *Parallel models of associative memory*. Lawrence Erlbaum. [JKF]
Holcomb, P. J. & Neville, H. J. (1991) Auditory and visual semantic priming in lexical decision: A comparison using event-related potentials. *Language & Cognitive Processes* 5:281–312. [VGH]
Horgan, J. (1996) *The end of science: Facing the limits of knowledge in the twilight of the scientific age.* Addison-Wesley. [rMJF]
Jacobs, R., Jordan, M. & Barto, A. (1990) Task decomposition through competition in a modular connectionist architecture: The What and Where vision tasks. *Technical Report* COINS TR-90–27, Department of Computer and Information Science, University of Massachusetts. [HB]
Jones, R. S. G. (1993) Entorhinal-hippocampal connections: A speculative view of their function. *Trends in Neurosciences* 16:58–64. [WEC]
Jordan, M. & Jacobs, R. (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2):181–214. [HB]
Kosslyn, S. M. & Intriligator, J. M. (1992) Is cognitive neuropsychology possible? The perils of sitting on a one-legged stool. *Journal of Cognitive Neuroscience* 4:96–106. [WEC]
Lakatos, I. (1974) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge,* ed. I. Lakatos & A. Musgrave. Cambridge University Press. [JKF]
Lipp, H. -P., Schwegler, H., Crusio, W. E., Wolfer, D., Leisinger-Trigona, M. -C., Heimrich, B. & Driscoll, P. (1989) Using genetically-defined rodent strains for the identification of hippocampal traits relevant for two-way avoidance learning: A non-invasive approach. *Experientia* 45:845–59. [WEC]
Logothetis, N. K., Pauls, J., Bulthoff, H. H. & Poggio, T. (1994) View-dependent object recognition by monkeys. *Current Biology* 4(5):401–413. [PJB]
McCarthy, R. A. & Warrington, E. K. (1990) *Cognitive neuropsychology: A clinical introduction*. Academic Press. [VG]
McClelland, J. L. & Rumelhart, D. E. (1986) A distributed model of human learning and memory. In: *Parallel distributed processing, vol. 2,* ed. J. L. McClelland & D. E. Rumelhart. MIT Press. [VG]
(1989) *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. MIT Press. [VG]
Neville, H. J. & Weber-Fox, C. (1994) Cerebral subsystems within language. In: *Structural and functional organization of neocortex: Proceedings of a symposium in memory of Otto D. Creutzfeldt, May 1993,* ed. B. Albowitz, K. Albus, V. Kuhnt, H-Ch. Nothdurft. & P. Wahle. Springer-Verlag. [VGH]
Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83. [VG]
Perrett, D. I., Hietanen, J. K., Oram, M. W. & Benson, P. J. (1992) Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London* B335(1273):23–30. [PJB]
Posner, M. I. (1994) Local and distributed processes in attentional orienting. *Behavioral and Brain Sciences* 17:78–79. [WEC]
Pribram, K. (1971) *Languages of the brain: Experimental paradoxes and principles in neuropsychology.* Prentice-Hall. [JKF]
(1982) Localization and distribution of function in the brain. In: *Neuropsychology after Lashley,* ed. J. Orbach. Lawrence Erlbaum. [JKF]
Pylyshyn, Z. W. (1984) *Computation and cognition: Toward a foundation for cognitive science.* A Bradford Book, The MIT Press. [VG]
Rumelhart, D. E. & McClelland, J. (1986) Parallel distributed processing, vols. I & II. MIT Press. [JKF]
Schoenfeld, T. A. & Hamilton, L. W. (1977) Secondary brain changes following lesions: A new paradigm for lesion experimentation. *Physiology and Behaviour* 18:951–67. [JKF]
Shallice, T. (1988) *From neuropsychology to mental structure.* Cambridge University Press. [VG]
Squire, L. (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99:195–231. [JKF]
Squire, L. R., Ojemann, J. G., Miezin, F. M., Petersen, S. S., Videen, T. O. & Raichle, M. E. (1992) Activation of the hippocampus in normal humans: A functional anatomical study of memory. *Proceedings of the National Academy of Sciences, USA* 89:1837–41. [WEC]
Tanaka, K., Saito, H.-A., Fukada, Y. & Moriya, M. (1991) Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology* 66:170–89. [PJB]
Tsotsos, J. (1990) Analyzing vision at the complexity level. *Behavioral and Brain Sciences* 13(3):423–69. [HB]
Vallar, G. (1991) Current methodological issues in human neuropsychology. In: *Handbook of neuropsychology,* vol. 5, ed. F. Boller & J. Grafman. Elsevier. [VG]
van Gelder, T. (1994) Playing Flourens to Fodor's Gall. *Behavioral and Brain Sciences* 17:84. [WEC]
Wahlsten, D. & Schalomon, P. M. (1994) A new hybrid mouse model for agenesis of the corpus callosum. *Behavioral Brain Research* 64:111–17. [WEC]
Walsh, V., Carden, D., Butler, S. R. & Kulikowski, J. J. (1993) The effects of V4 lesions on the visual abilities of Macaques: Hue discrimination and colour constancy. *Behavioral Brain Research* 53(1–2):51–62. [PJB]
Warrington, E. K. & Sallice, T. (1984) Category specific semantic impairments. *Brain* 107:829–54. [WEC]
Webster, W. G. (1973) Assumptions, conceptualizations, and the search for the functions of the brain. *Physiological Psychology* 1:346–50. [JKF]
Weiskrantz, L. (1968) Treatments, inferences, and brain function. In: *Analysis of behavioural change,* ed. L. Weiskrantz. Harper & Row. [JKF]
(1974) Brain research and parallel processing. *Physiological Psychology* 2:53–54. [JKF]
Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M. & Hay, D. C. (1993) Face perception after brain injury: Selective impairments affecting identity and expression. *Brain* 116:941–59. [PJB]