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-ouput 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 backpropagation 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 processes 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 backpropagation. 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 SPG-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 perforce 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-