Modularity, interaction and connectionist neuropsychology

Nick Chater

Neural Networks Research Group, Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom
Electronic mail: *nicholas@cogsci.ed.ac.uk*

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

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

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

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

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

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

ACKNOWLEDGMENT

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