Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press. Pp. 208. ISBN 0–262–13379–2. £18.50/$27.95 (hardback).

In a critique of a nascent connectionist cognitive science, Fodor and Pylyshyn (1988) issued a dilemma to researchers developing neural network models of cognition. They argued that connectionist models either merely implement symbolic systems, or fail to capture essential properties of human cognition. This dilemma depends upon properties that they deemed requisite to cognition. Gary Marcus adapts the same style of argumentation in his recent book *The Algebraic Mind*, without the associated pessimism for symbolic implementation. Marcus' book is organised according to three important features of our cognitive architecture as he sees it. Any model of human cognition, he suggests, must allow abstract relations between variables and recursive representations, along with an ability to distinguish between kinds and individuals. Much of the book is devoted to demonstrating that certain implementations of popular types of connectionist models, multilayer perceptrons and simple recurrent networks, cannot account for these aspects of cognition.

The first two properties Marcus proposes, abstract relations and recursive representation, can be traced back at least to Fodor's (1975) ''language of thought'' hypothesis. By abstract relations, Marcus means open-ended schemas or rules that hold between whole classes of entities, much like algebraic equations in mathematics or computer programming (p. 35). He asserts that the mind represents these relations and other facts recursively in that new knowledge can be constructed by combining simpler elements into more complex elements. Both of these properties underlie a kind of language of thought or ''mentalese'' described by Fodor long ago. In fact, Fodor and Pylyshyn (1988) maintained that the systematicity inherent in mentalese cannot be captured by connectionist models, at least not without implementing a symbolic cognitive system. Marcus appears to adapt a similar approach in this book.

A third essential property Marcus discusses, the ability to distinguish between kinds and individuals, is a novel approach to criticising connectionist models. Marcus' coverage of the relevant literature, however, leaves much to be desired. It is unclear at this time that such a skill is as widespread in our cognitive system as he suggests. Indeed, research in social cognition, for example, reveals a blurring of kinds and individuals at some levels of processing (Banaji, Hardin, & Rothman, 1993). A wider canvassing of the relevant empirical literature is needed before claiming that this skill is as much of a desideratum for models as Marcus argues.

As case examples, Marcus considers several neural network implementations that fail to satisfy the first two properties of abstract relations and recursive representation. These limitations stem from well-known properties of network dynamics, and are usually overcome in practice by choosing input representa-

tions and training regimes suitable to the task at hand. For example, Marcus describes a multilayer perceptron of his own that copies or inverts a binary number (p. 49). His model failed to copy sequences not seen in training. However, our own exploratory simulations revealed that sensible changes to Marcus' design (e.g., total connectivity and some additional copy training) will result in the expected performance. He also demonstrates that some prominent models fail to make the distinction between kinds and individuals, though it should be noted that these models were actually not originally developed for making such distinctions. Ironically, Marcus seems to mirror the failings of his own networks on a theoretical level when he argues from the failure of a few individual models that the whole kind of ''multilayer perceptrons do not offer an adequate basis for cognition'' (p. 7). But if there is one area of human endeavour in which distinguishing between kinds and individuals is crucial, it is in arguments regarding the relative merits of scientific theories. Thus, in modern post-Popperian philosophy of science, disconfirming individual instantiations of a theory does not necessarily falsify the theory as a whole. We therefore urge the reader not to confuse problems with individual network implementations for problems with kinds of networks as a whole.

The book's vision of our cognitive architecture and its critical approach to connectionism are heavily dependent upon a Fodorian foundation. The three properties of cognition that organise the core of the book have a long pedigree in symbolic theory. In a later chapter on evolutionary psychology, he seems to embrace the modularity hypothesis that often marks similar perspectives (pp. 146, 150). The overall enthusiasm for symbol manipulation and modularity surfaces clearly in the final pages of the book: ''To understand human cognition, we need to understand how basic computational components are integrated into more complex devices—such as parsers, language acquisition devices, modules for recognizing objects, and so forth'' (p. 172). However, there may be reason to question that this framework can offer as deep an understanding as Marcus hopes. Fodor, the framework's forefather, has himself recently issued arguments that should temper such optimism. Discussing classical computation, modularity, and adaptationism, Fodor writes: ''The three together constitute not an utterly implausible account of some aspects of cognition. As the reader will no doubt have noticed, it's the part of cognition that *doesn't* work that way that I'm worried about, the indications being that it's quite a big part, and that much of what's special about our kinds of minds lives there'' (Fodor, 2000, p. 80). Fodor argues that there are vast regions of our cognitive architecture about which Marcus' approach would have nothing to say, and that it is ''light years from being satisfactory'' (Fodor, 2000, p. 5). *The Algebraic Mind* therefore seems unbalanced in its criticisms, offering only a unidirectional critique of eliminative connectionism. Perhaps more revealing, the book offers little in terms of what actual symbolic implementations could replace the ones he criticises (for example, even though he offers fairly detailed suggestions for a theory of

‘‘treelets’’ [p. 108], they are not accompanied by any implementation). One cannot help wonder why Marcus devotes so much effort to connectionist implementations that do not work rather than offering up algebraic ones that do.

Marcus’ book incorporates lines of criticism against neural network models that are instructional, and lead to important discussion and debate. In his own demonstrations of network failings, he offers lessons on what to avoid and overcome when designing connectionist models. The book does suffer from only considering potential problems of eliminative connectionism, without equal weight dedicated to discussing the well-known shortcomings of the symbolic approach; and this despite a subtitle that promises to *integrate* connectionism and cognitive science. A more suitable subtitle, one that would at least lead to appropriate expectations for the reader, might have been ‘‘Assimilating Connectionism into Symbolic Cognitive Science’’.

REFERENCES

Banaji, M. R., Hardin, C., & Rothman, A. J. (1993). Implicit stereotyping in person judgment. *Journal of Personality and Social Psychology*, *65*, 272–281.

Fodor, J. (1975). *The language of thought*. New York: Crowell.

Fodor, J. (2000). *The mind doesn’t work that way*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *78*, 3–71.

RICK DALE AND MORTEN H. CHRISTIANSEN
Cornell University, Ithaca, USA