Chapter 1
Cognitive Architecture

What is a mind such that it can entertain an infinity of thoughts? Is it a manipulator of symbols, as the late Allen Newell (1980) suggested? Or is it a device in which "the basic unit[s] of cognition" have nothing "essential to do with sentences and propositions" of symbol-manipulation, as Paul Churchland (1995, p. 322) has suggested? In the last decade or so, this question has been one of the central controversies in cognitive science. Interest in this question has largely been driven by a set of researchers who have proposed *neural network* or *connectionist* models of language and cognition. Whereas *symbol-manipulating* models are typically described in terms of elements like *production rules* (if preconditions 1, 2, and 3 are met, take actions 1 and 2) and *hierarchical binary trees* (such as might be found in a linguistics textbook), connectionist models are typically meant to be "neurally-inspired" and are typically described in terms of basic elements such as neuronlike nodes and synapse-like *connections*. Such models are sometimes said not to "look like anything we have ever seen before" (Bates & Elman, 1993, p. 637), and for this reason, connectionist models have sometimes been described as signaling a *paradigm shift* in cognitive science (Bechtel & Abrahamsen, 1991; Sampson, 1987; Schneider, 1987).

But surface appearances can be deceiving. As it turns out, some models can be both connectionist and symbol-manipulating at the same time. For example, symbol-manipulating models standardly make use of logical functions like AND and OR, and it turns out those functions can easily be built in—or, *implemented in*—connectionist nodes. In fact, perhaps the first discussion about how cognition might be implemented in neural substrate was a discussion by McCulloch and Pitts (1943) of how "a logical calculus [of] ideas"—functions like AND and OR—could be built of neuronlike nodes.[1]

The mere fact that the brain is made up (in large part) of neurons does not by itself tell us whether the brain implements the machinery of symbol-manipulation (rules and the like). Instead, the question of whether the brain implements the machinery of symbol-manipulation is

a question about how basic computational units are put together into more complex circuits. Advocates of symbol-manipulation assume that the circuits of the brain correspond in some way to the basic devices assumed in discussions of symbol-manipulation—for example, that some kind of brain circuit that supports the representation (or generalization) of a rule. Critics of symbol-manipulation argue that there will not turn out to be brain circuits that implement rules and the like.

In keeping with this basic tension, the term *connectionism* turns out to be ambiguous. Most people associate the term with the researchers who have most directly challenged the symbol-manipulation hypothesis, but the field of connectionism also encompasses models that have sought to explain how symbol-manipulation can be implemented in a neural substrate (e.g., Barnden, 1992b; Hinton, 1990; Holyoak, 1991; Holyoak & Hummel, 2000; Lebière & Anderson, 1993; Touretzky & Hinton, 1985).

This systematic ambiguity in what is meant by the term *connectionism* has, in my view, impaired our understanding of the relation between connectionism and symbol-manipulation. The problem is that discussions of the relation between connectionism and symbol-manipulation often assume that evidence *for* connectionism automatically counts as evidence *against* symbol-manipulation. But because connectionist models vary widely in their architectural and representational assumptions, collapsing them together can only obscure our understanding of the relation between connectionism and symbol-manipulation.

The burden of proof in understanding the relation between connectionism and symbol-manipulation should be shared equally. There is no default about whether a given connectionist model implements a particular aspect of symbol-manipulation: some models will, some models will not. Deciding whether a given model implements symbol-manipulation is an empirical question for investigation and analysis that requires a clear understanding of symbol-manipulation and a clear understanding of the model in question. Only with an understanding of both can we tell whether that model offers a genuine alternative to Newell's position that the mind is a manipulator of symbols.

1.1 Preview

My aim in this book is to integrate the research on connectionist models with a clear statement about what symbol-manipulation is. My hope is that we can advance beyond earlier discussions about connectionism and symbol-manipulation by paying special attention to the differences between different connectionist models and to the relationship between particular models and the particular assumptions of symbol-manipulation.

I do not cast the debate in quite the terms that it has been cast before. For one thing, I do not adopt Pinker and Prince's (1988) distinction between *eliminative connectionism* and *implementational connectionism.* Although I have used these terms before, I avoid them here for several reasons. First, people often associate the word "mere" with implementational connectionism, as if implementational connectionism were somehow an unimportant research project. I avoid such negative connotations because I strongly disagree with their premise. If it turns out that the brain does in fact implement symbol-manipulation, implementational connectionism would be far from unimportant. Instead, it would be an enormous advance, tantamount to figuring out how an important part of the brain really works. Second, although many researchers have challenged the idea of symbol-manipulation, few self-identify as advocates of eliminative connectionism. Instead, those who have challenged symbol-manipulation typically self-identify as connectionists without explicitly specifying what version of connectionism they favor. The consequence is that it is hard to point to clear statements about what eliminative connectionism is (and it is also hard to discern the relation between particular models and the hypotheses of symbol-manipulation). Rather than focusing on such an ill-defined position, I instead focus on a particular class of models—*multilayer perceptrons.* My focus is on these models because these are almost invariably the ones being discussed when researchers consider the relation between connectionism and symbol-manipulation. Part of the work to be done is to carefully specify the relation between those models and the hypothesis of symbol-manipulation. To assume in advance that multilayer perceptrons are completely inconsistent with symbol-manipulation would be to unfairly prejudge the issue.

Another way in which my presentation will differ is that in contrast to some other researchers, I couch the debate not as being about symbols but as being about symbol-*manipulation.* In my view, it is simply not useful to worry about whether multilayer perceptrons make use of symbols *per se.* As far I can tell (see section 2.5), that is simply a matter of definitions. The real work in deciding between competing accounts of cognitive architecture lies not in what we call symbols but in understanding what sorts of representations are available and what we do with them.

In this connection, let me stress that symbol-manipulation is not a single hypothesis but a family of hypotheses. As I reconstruct it, symbol-manipulation consists of three separable hypotheses:

- The mind represents *abstract relationships* between *variables.*
- The mind has a system of *recursively structured representations.*

- The mind distinguishes between mental representations of *individuals* and mental representations of *kinds*.

I detail what I mean by these hypotheses later. For now, my point is only that these hypotheses can stand or fall separately. It could turn out that the mind makes use of, say, abstract representations of relationships between variables but does not represent recursively structured knowledge and does not distinguish between mental representations of individuals and mental representations of kinds. Any given model, in other words, can be consistent with one subset of the three hypotheses about symbol-manipulation or with all of them. A simple dichotomy between implementational connectionism and eliminative connectionism does not capture this.

I therefore instead evaluate each of the hypotheses of symbol-manipulation separately. In each case I present a given hypothesis and ask whether multilayer perceptrons offer alternatives to it. Where multilayer perceptrons do offer an alternative, I evaluate that alternative. In all cases, I suggest accounts of how various aspects of mental life can be implemented in neural machinery.

Ultimately, I argue that models of language and cognition that are consistent with the assumptions of symbol-manipulation are more likely to be successful than models that are not. The aspects of symbol-manipulation that I defend—symbols, rules, variables, structured representations, and distinct representations of individuals—are not new. J. R. Anderson, for example, has through the years adopted all of them in his various proposals for cognitive architecture (e.g., Anderson, 1976, 1983, 1993). But we are now, I believe, in a better position to evaluate these hypotheses. For example, writing prior to all the recent research in connectionism, Anderson (1976, p. 534) worried that the architecture that he was then defending might "be so flexible that it really does not contain any empirical claims and really only provides a medium for psychological modeling." But things have changed. If in 1976 Anderson had little to use as a point of comparison, the advent of apparently paradigm-shifting connectionist models now allows us to see that assumptions about symbol-manipulation are falsifiable. There are genuinely different ways in which one might imagine constructing a mind.[2]

The rest of this book is structured as follows. Chapter 2 is devoted to explaining how multilayer perceptrons work. Although these are not the only kind of connectionist models that have been proposed, they deserve special attention, both because they are the most popular and because they come closer than any other models to offering a genuine, worked-out alternative to symbol-manipulation.

In chapters 3, 4, and 5, I discuss what I take to be the three core tenets of symbol-manipulation, in each case contrasting them with the as-

sumptions implicit in multilayer perceptron approaches to cognition. Chapter 3 considers the claim that the mind has mechanisms and representational formats that allow it to represent, extract, and generalize abstract relationships between mentally represented variables—relationships that sometimes are known as *rules*.[3] These entities would allow us to learn and represent relationships that hold for all members of some of class, and to express generalizations compactly (Barnden, 1992a; Kirsh, 1987). Rather than specifying individually that *Daffy likes to swim, Donald likes to swim,* and so forth, we can describe a generalization that does not make reference to any specific duck, thereby using the type **duck** as an implicit variable. In this way, variables act as placeholders for arbitrary members of a category.

Going somewhat against the conventional wisdom, I suggest that multilayer perceptrons and rules are not entirely in opposition. Instead, the real situation is more subtle. All multilayer perceptrons can in principle represent abstract relationships between mentally represented variables, but only some actually do so. Furthermore, some—but not all—can acquire rules on the basis of limited training data. In a pair of case studies, I argue that the only models that adequately capture certain empirical facts are those that implement abstract relations between variables.

Chapter 4 defends the claim that the mind has ways of internally representing structured knowledge—distinguishing, for example, between mental representations of *the book that is on the table* and mental representations of *the table that is on the book.* I show that the representational schemes most widely used in multilayer perceptrons cannot support such structured knowledge but suggest a novel account for how such knowledge could be implemented in a neural substrate.

Chapter 5 defends the claim that the mind represents a distinction between kinds and individuals—distinguishing, for example, between Felix and cats in general. I show that, in contrast, the representational schemes most widely used in multilayer perceptrons cannot support a distinction between kinds and individuals. The chapter ends with some brief remarks about how such a distinction could be implemented.

Following these chapters, I provisionally accept the hypothesis that the mind manipulates symbols, and in chapter 6 take up the questions of how the machinery for symbol-manipulation could develop in the mind of the child and how that machinery could have been shaped across evolutionary time. Chapter 7 concludes.

Throughout this book, I use the following notational conventions: **boldface** for variables and nodes; *italics* for words that are mentioned rather than used; SMALL CAPS for mental representations of kinds (cats, dogs,

and so forth). Thus the concept of a cat would be represented internally by the kind CAT, represented in a neural network by a node called **cat,** and represented in English by the word *cat*.

1.2 Disclaimers

In keeping with a point that I stressed in the preface, let me again emphasize that I do not argue that no form of connectionism can succeed. Rather, I am laying out a geography of possible models and making suggestions about which I think are most likely to succeed.

I close this introduction with two caveats. First, my empirical focus is on language and higher-level cognition rather than, say, perception and action partly because language and cognition are the domains that I am most familiar with and partly because these are the domains most often described in terms of symbol-manipulation. If symbol-manipulation does not play a role in language and higher-level cognition, it seems unlikely that it plays a role in other domains. Of course, the reverse is not true; it is perfectly possible that symbol-manipulation plays a role in language and cognition without playing a role elsewhere. Rather than trying to settle these issues about other domains here, my hope is that the discussion I present can serve as a guide to those who want to investigate analogous questions about the role of symbol-manipulation in other domains.

As a second caveat, to the extent that part of this book serves as a critique, it must serve as a critique of multilayer perceptrons and not as a critique of possible alternatives to symbol-manipulation that have not yet been proposed. In presenting this material, I have often encountered audiences that seem to want me to *prove* that the mind manipulates symbols. Of course, I can do no such thing. At most, I can show that symbol-manipulation is consistent with the facts and that the alternatives thus far proposed are inadequate. I cannot possibly rule out alternatives that have not yet been proposed. The situation here is the same as elsewhere in science: disconfirmation can be decisive, but confirmation is just an invitation for further investigation.