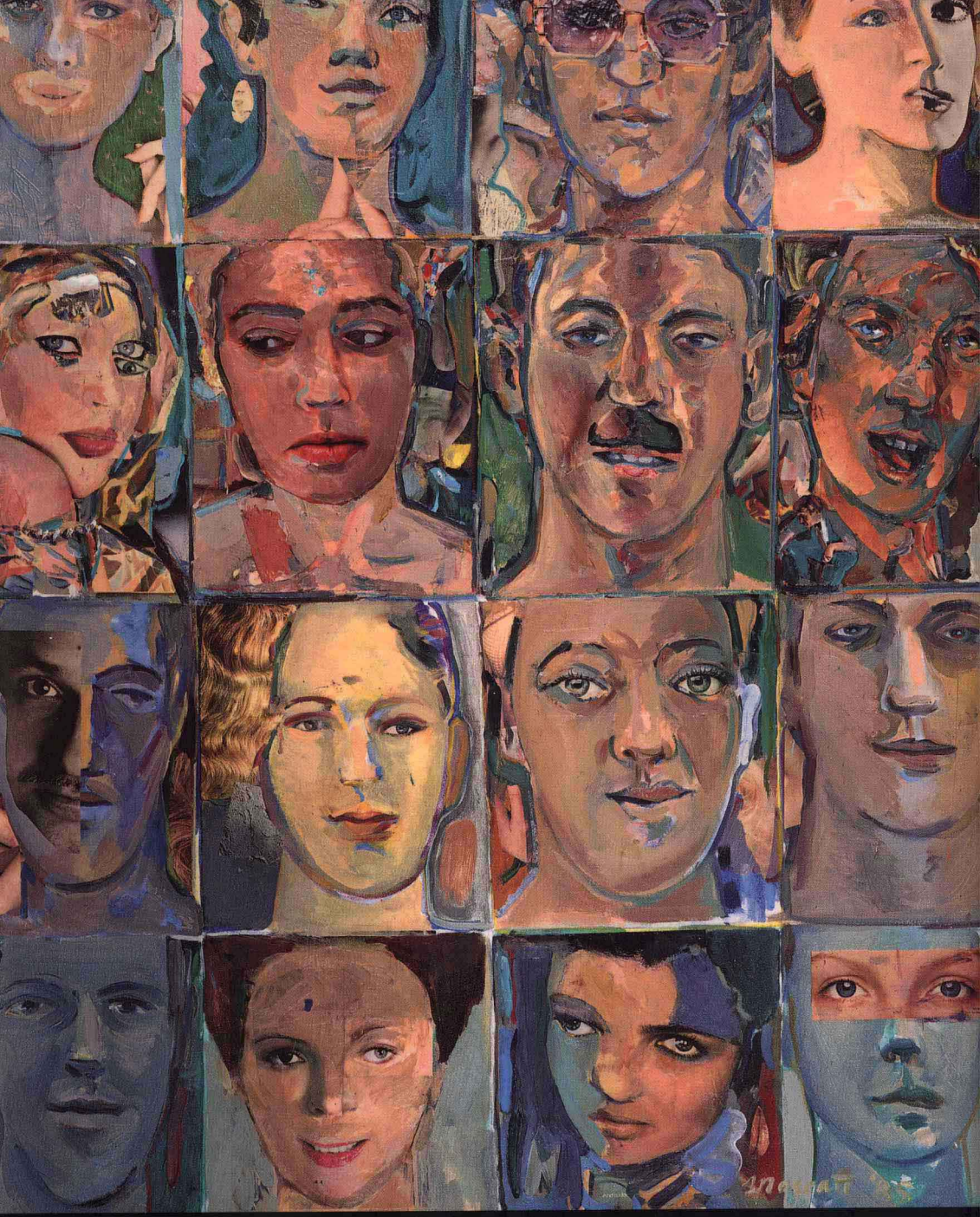

The Algebraic Mind

Integrating Connectionism and Cognitive Science

Gary F. Marcus

The Algebraic Mind

Integrating Connectionism and Cognitive Science

Gary F. Marcus

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

© 2001 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Palatino by Graphic Composition, Inc., and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Marcus, Gary F. (Gary Fred)
The algebraic mind: integrating connectionism and cognitive science / Gary F. Marcus.
p. cm. — (Learning, development, and conceptual change)
"A Bradford book."
Includes bibliographical references and index.
ISBN 0-262-13379-2 (hard: alk. paper)
1. Mental representation. 2. Cognition. 3. Connectionism. 4. Cognitive science. I. Title. II. Series.

BF316.6 .M35 2001
153—dc21

00-038029

The Algebraic Mind

LD&CC Learning, Development, and Conceptual Change
Lila Gleitman, Susan Carey, Elissa Newport, and Elizabeth Spelke, editors

From Simple Input to Complex Grammar, James L. Morgan, 1986

Concepts, Kinds, and Cognitive Development, Frank C. Keil, 1989

Learnability and Cognition: The Acquisition of Argument Structure, Steven Pinker, 1989

Mind Bugs: The Origins of Procedural Misconception, Kurt VanLehn, 1990

Categorization and Naming in Children: Problems of Induction, Ellen M. Markman, 1989

The Child's Theory of Mind, Henry M. Wellman, 1990

Understanding the Representational Mind, Josef Perner, 1991

An Odyssey in Learning and Perception, Eleanor J. Gibson, 1991

Beyond Modularity: A Developmental Perspective on Cognitive Science, Annette Karmiloff-Smith, 1992

Mindblindness: An Essay on Autism and "Theory of Mind," Simon Baron-Cohen, 1995

Speech: A Special Code, Alvin M. Liberman, 1995

Theory and Evidence: The Development of Scientific Reasoning, Barbara Koslowski, 1995

Race in the Making: Cognition, Culture, and the Child's Construction of Human Kinds, Lawrence A. Hirschfeld, 1996

Words, Thoughts, and Theories, Alison Gopnik and Andrew N. Meltzoff, 1996

The Cradle of Knowledge: Development of Perception in Infancy, Philip J. Kellman and Martha E. Arterberry, 1998

Language Creation and Change: Creolization, Diachrony, and Development, edited by Michel DeGraff, 1999

Systems That Learn: An Introduction to Learning Theory, second edition, Sanjay Jain, Daniel Osherson, James S. Royer, and Arun Sharma, 1999

How Children Learn the Meanings of Words, Paul Bloom, 2000

Making Space: The Development of Spatial Representation and Reasoning, Nora S. Newcombe and Janellen Huttenlocher, 2000

The Algebraic Mind: Integrating Connectionism and Cognitive Science, Gary F. Marcus, 2001

此为试读,需要完整PDF请访问: www.ertongbook.com

Thought is in fact a kind of algebra, as Berkeley long ago said, "in which, though a particular quantity be marked by each letter, yet to proceed right, it is not requisite that in every step each letter suggest to your thoughts that particular quantity it was appointed to stand for."
—William James, *Principles of Psychology*

[I have] sometimes heard it said that the nervous system consists of huge numbers of random connections. Although its orderliness is indeed not always obvious, I nevertheless suspect that those who speak of random networks in the nervous system are not constrained by any previous exposure to neuroanatomy.
—David Hubel, *Eye, Brain, and Vision*

Series Foreword

This series in learning, development, and conceptual change includes state-of-the-art reference works, seminal book-length monographs, and texts on the development of concepts and mental structures. It spans learning in all domains of knowledge, from syntax to geometry to the social world, and is concerned with all phases of development, from infancy through adulthood.

The series intends to engage such fundamental questions as the following:

The nature and limits of learning and maturation The influence of the environment, of initial structures, and of maturational changes in the nervous system on human development; learnability theory; the problem of induction; and domain-specific constraints on development

The nature of conceptual change Conceptual organization and conceptual change in child development, in the acquisition of expertise, and in the history of science

Lila Gleitman
Susan Carey
Elissa Newport
Elizabeth Spelke

Preface

My interest in cognitive science began in high school, with a naïve attempt to write a computer program that I hoped would translate Latin into English. The program didn't wind up being able to do all that much, but it brought me to read some of the literature on artificial intelligence. At the center of this literature was the metaphor of mind as machine.

Around the time that I started college, cognitive science had begun an enormous shift. In a two-volume book called *Parallel Distributed Processing* (or just PDP), David E. Rumelhart, James L. McClelland, and their collaborators (McClelland, Rumelhart & the PDP Research Group, 1986; Rumelhart, McClelland & the PDP Research Group, 1986) argued that the mind was not nearly as much like a computer as I had once thought. Instead, these researchers favored what they called *neural networks* or *connectionist models.* I was hooked immediately and thrilled when I managed to find a summer job doing some PDP-like modeling of human memory. Although my undergraduate thesis was not about PDP models (it was instead about human reasoning), I never lost interest in questions about computational modeling and cognitive architecture.

When I was searching for graduate programs, I attended a brilliant lecture by Steven Pinker in which he compared PDP and symbol-manipulation accounts of the inflection of the English past tense. The lecture convinced me that I needed to work with Pinker at MIT. Soon after I arrived, Pinker and I began collaborating on a study of children's overregularization errors (*breaked, eated,* and the like). Infected by Pinker's enthusiasm, the minutiae of English irregular verbs came to pervade my every thought.

Among other things, the results we found argued against a particular kind of neural network model. As I began giving lectures on our results, I discovered a communication problem. No matter what I said, people would take me as arguing against *all* forms of connectionism. No matter how much I stressed the fact that other, more sophisticated kinds of network models were left untouched by our research, people always seem to come away thinking, "Marcus is an anti-connectionist."

But I am not an anti-connectionist; I am opposed only to a particular *subset* of the possible connectionist models. The problem is that the term *connectionism* has become synonymous with a single kind of network model, a kind of empiricist model with very little innate structure, a type of model that uses a learning algorithm known as *back-propagation*. These are not the only kinds of connectionist models that *could* be built; indeed, they are not even the only kinds of connectionist models that *are* being built, but because they are so radical, they continue to attract most of the attention.

A major goal of this book is to convince you, the reader, that the type of network that gets so much attention occupies just a small corner in a vast space of possible network models. I suggest that adequate models of cognition most likely lie in a different, less explored part of the space of possible models. Whether or not you agree with my specific proposals, I hope that you will at least see the value of exploring a broader range of possible models. Connectionism need not just be about back-propagation and empiricism. Taken more broadly, it could well help us answer the twin questions of what the mind's basic building blocks are and how those building blocks can be implemented in the brain.

All the mistakes in this book are my own, but much of what is right I owe to my colleagues. My largest and most obvious debt is to Steve Pinker, for the excellent training he gave me and for the encouragement and meticulous, thought-provoking comments that he continues to supply. I owe similar debts to my undergraduate advisors Neil Stillings and Jay Garfield, each of whom spent many hours teaching me in my undergraduate years at Hampshire College and each of whom provided outstanding comments on earlier drafts of this book.

Going back even further, my first teacher was my father, Phil Marcus. Although technically speaking he is not a colleague, he frequently asked me important theoretical questions that helped me to clarify my thoughts.

Susan Carey has been an unofficial mentor to me since I arrived at NYU. For that, and for incisive comments that helped me turn a rough draft into a final draft, I am very grateful.

A great many other colleagues provided enormously helpful, detailed comments on earlier drafts of this book, including Iris Berent, Paul Bloom, Luca Bonatti, Chuck Clifton, Jay Garfield, Peter Gordon, Justin Halberda, Ray Jackendoff, Ken Livingston, Art Markman, John Morton, Mike Nitabach, Michael Spivey, Arnold Trehub, Virginia Valian, and Zsófia Zvolenszky. Ned Block, Tecumseh Fitch, Cristina Sorrentino, Travis Williams, and Fei Xu each made trenchant comments on particular chapters. For their helpful discussion and patient answers to my queries, I would also like to thank Benjamin Bly, Noam Chomsky, Har-

ald Clahsen, Dan Dennett, Jeff Elman, Jerry Fodor, Randy Gallistel, Bob Hadley, Stephen Hanson, Todd Holmes, Keith Holyoak, John Hummel, Mark Johnson, Denis Mareschal, Brian McElree, Yuko Munakata, Mechiro Negishi, Randall O'Reilly, Neal Perlmutter, Nava Rubin, Lokendra Shastri, Paul Smolensky, Liz Spelke, Ed Stein, Wendy Suzuki, Heather van der Lely, and Sandy Waxman, and my colleagues at UMass/Amherst (where I began this project) and NYU (where I finished it). I also thank my research assistants, Shoba Bandi Rao and Keith Fernandes, for their help in running my lab and all the students who took my spring 1999 graduate course on computational models of cognitive science. I'd like to thank MIT Press, especially Amy Brand, Tom Stone, and Deborah Cantor-Adams, for their help in producing the book. NIH Grant HD37059 supported the final stages of the preparation of this book.

My mother, Molly, may not share my interest in irregular verbs or neural networks, but she has long encouraged my intellectual curiosity. Both she and my friends, especially Tim, Zach, Todd, Neal, and Ed, have helped me to maintain my sanity throughout this project.

Finally, and not just because she is last in alphabetical order, I wish to thank Zsófia Zvolenszky. Through her comments and her love, she has helped to make this book much better and this author much happier.

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

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,

此为试读,需要完整PDF请访问: www.ertongbook.com