The Cognitive Science of Science

Explanation, Discovery, and Conceptual Change

Paul Thagard

The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change

Paul Thagard
in collaboration with Scott Findlay, Abninder Litt, Daniel Saunders, Terrence C. Stewart, and Jing Zhu

The MIT Press
Cambridge, Massachusetts
London, England

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Stone Sans and Stone Serif by Toppan Best-set Premedia Limited. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Thagard, Paul.
The cognitive science of science : explanation, discovery, and conceptual change / Paul Thagard ; in collaboration with Scott Findlay . . . [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-01728-2 (hardcover : alk. paper)
1. Science—Philosophy. 2. Cognitive science. I. Findlay, Scott. II. Title.
Q175.T478 2012
501—dc23

2011039760

10 9 8 7 6 5 4 3 2 1

Contents

Preface

This book is a collection of my recent essays on the cognitive science of science that illustrate ways of combining philosophical, historical, psychological, computational, and neuroscientific approaches to explaining scientific development. Most of the chapters have been or will be published elsewhere, but the introductions are brand new (chapters 1, 2, 7, 12), as are the last two chapters, which take the cognitive science of science in new directions related to values and concepts. The reprinted chapters reproduce the relevant articles largely intact, but I have done some light editing to coordinate references and remove redundancies. Origins of the articles and coauthors are indicated in the acknowledgments.

Early in my career, I wandered into the cognitive science of science through a series of educational accidents, and have enthusiastically pursued research that is variously philosophical, historical, psychological, computational, and neurobiological. In high school, I did very well in physics and chemistry, but only because I was adept at solving math problems, not because I found science very interesting. As an undergraduate at the University of Saskatchewan, I avoided serious science courses, although I did get a good sampling of mathematics and logic. My interest in science was sparked during my second undergraduate degree at Cambridge University, where the philosophy course I took required a paper in philosophy of science. Through lectures by Ian Hacking and Gerd Buchdahl, along with books by Russell Hanson and Thomas Kuhn, I started to appreciate the value of understanding the nature of knowledge by attention to the history of science. I was struck by how much more rich and interesting the scientific examples of knowledge were compared to the contrived thought experiments favored by epistemologists working in the tradition of analytic philosophy. Accordingly, my Ph.D. work at the

University of Toronto focused on scientific reasoning enriched by historical case studies.

My move into cognitive science was also serendipitous. In 1978, in the second term of my first teaching job at the University of Michigan–Dearborn, I decided to sit in on a graduate epistemology course taught by Alvin Goldman at the main Michigan campus in Ann Arbor. It turned out that this course was coordinated with one on human inference taught by the social psychologist Richard Nisbett. The combination of these two courses was amazing: Goldman was pioneering an approach to epistemology that took experimental research on psychology seriously, and Nisbett was presenting a draft of his path-breaking book with Lee Ross, *Human Inference*. I started reading avidly in cognitive psychology, which quickly led me to the field of artificial intelligence. I was attracted by the theoretical ideas of visionaries such as Marvin Minsky, and also by the prospects of a new methodology—computer modeling—for understanding the structure and growth of scientific knowledge.

Accordingly, I did an MS in computer science at Michigan and started building my own computational models of various aspects of scientific thinking. My early models of analogical thinking were somewhat crude, but became much more powerful when my collaborator Keith Holyoak came up with the idea of modeling analogy using connectionist ideas about parallel constraint satisfaction. I quickly realized that theory choice based on the explanatory power of competing theories could also be simulated using neural networks.

My interest in neuroscience also came about indirectly. After I moved to Waterloo in 1992, one of my graduate students, Allison Barnes, was investigating empathy as a kind of analogy, which led me to general concern with emotions. The work of Antonio Damasio revealed how crucial neuroscience was to understanding emotions, and since I was already building artificial neural network models, it was natural to try to undertake more realistic neural models of emotion and decision making. Happily, this line of work has turned back around to scientific applications, described in some of the chapters below.

I remain convinced that understanding the growth of knowledge requires the kind of interdisciplinary approach found in cognitive science. I hope this collection will appeal to anyone interested in the structure and growth of scientific knowledge, including scientists, philosophers, historians, psychologists, sociologists, and educators.

Acknowledgments

While I was writing and revising this work, my research was supported by the Natural Sciences and Engineering Research Council of Canada. I am grateful to the coauthors of essays included in this collection: Scott Findlay (who made major contributions to three chapters), Abninder Litt, Daniel Saunders, Terry Stewart, and Jing Zhu. Please note that Daniel is first author of our joint article. Chris Eliasmith's exciting ideas about theoretical neuroscience contributed to several chapters. For comments or suggestions for particular chapters, I am indebted to him, William Bechtel, Chris Grisdale, Lloyd Elliott, Robert Hadley, Phil Johnson-Laird, Kostas Kampourakis, Eric Lormand, Elijah Millgram, Daniel Moerman, Nancy Nersessian, Eric Olsson, Robert Proctor, Peter Railton, David Rudge, Daniel Saunders, Cameron Shelley, and Terry Stewart. CBC Radio 2 provided the accompaniment.

I am grateful to my coauthors and to the respective publishers for permission to reprint the following essays:

Thagard, P., & Litt, A. (2008). Models of scientific explanation. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 549–564). Cambridge: Cambridge University Press. © Cambridge University Press.

Thagard, P. (2010). How brains make mental models. In L. Magnani, W. Carnielli & C. Pizzi (Eds.), *Model-based reasoning in science and technology: Abduction, logic, and computational discovery* (pp. 447–461). Berlin: Springer. © Springer.

Thagard, P., & Findlay, S. D. (2011). Changing minds about climate change: Belief revision, coherence, and emotion. In E. J. Olsson & S. Enqvist (Eds.), *Belief revision meets philosophy of science* (pp. 329–345). Berlin: Springer. © Springer.

Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science, 74,* 28–47. © University of Chicago Press.

Thagard, P., & Stewart, T. C. (2011). The Aha! experience: Creativity through emergent binding in neural networks. *Cognitive Science, 35,* 1–33. © Cognitive Science Society.

Thagard, P. (forthcoming). Creative combination of representations: Scientific discovery and technological invention. In R. Proctor & E. J. Capaldi (Eds.), *Psychology of science*. Oxford: Oxford University Press. © Oxford University Press.

Saunders, D., & Thagard, P. (2005). Creativity in computer science. In J. C. Kaufman & J. Baer (Eds.), *Creativity across domains: Faces of the muse* (pp. 153–167). Mahwah, NJ: Lawrence Erlbaum. © Taylor and Francis.

Thagard, P. (2011). Patterns of medical discovery. In F. Gifford (Ed.), *Handbook of philosophy of medicine* (pp. 187–202). Amsterdam: Elsevier. © Elsevier.

Thagard, P. (2008). Conceptual change in the history of science: Life, mind, and disease. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 374–387). London: Routledge. © Taylor and Francis.

Thagard, P., & Findlay, S. (2010). Getting to Darwin: Obstacles to accepting evolution by natural selection. *Science & Education*, 19, 625–636. © Springer.

Thagard, P., & Zhu, J. (2003). Acupuncture, incommensurability, and conceptual change. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 79–102). Mahwah, NJ: Lawrence Erlbaum. © Taylor and Francis.

Thagard, P., & Findlay, S. (forthcoming). Conceptual change in medicine: Explanations of mental illness from demons to epigenetics. In W. J. Gonzalez (Ed.), *Conceptual revolutions: From cognitive science to medicine*. A Coruña, Spain: Netbiblo. © Netbiblo.

Finally, I am grateful to Judith Feldmann for skillful editing and to Eric Hochstein for help with the index.

I Introduction

1 What Is the Cognitive Science of Science?

Explaining Science

Science is one of the greatest achievements of human civilization, contributing both to the acquisition of knowledge and to people's well-being through technological advances in areas from medicine to electronics. Without science, we would lack understanding of planetary motion, chemical reactions, animal evolution, infectious disease, mental illness, social change, and countless other phenomena of great theoretical and practical importance. We would also lack many valuable applications of scientific knowledge, including antibiotics, airplanes, and computers. Hence it is appropriate that many disciplines such as philosophy, history, and sociology have attempted to make sense of how science works.

This book endeavors to understand scientific development from the perspective of cognitive science, the interdisciplinary investigation of mind and intelligence. Cognitive science encompasses at least six fields: psychology, neuroscience, linguistics, anthropology, philosophy, and artificial intelligence (for overviews, see Bermudez, 2010; Gardner, 1985; Thagard, 2005a). The main intellectual origins of cognitive science are in the 1950s, when thinkers such as Noam Chomsky, George Miller, Marvin Minsky, Allan Newell, and Herbert Simon began to develop new ideas about how human minds and computer programs might be capable of intelligent functions such as problem solving, language, and learning. The organizational origins of cognitive science are in the 1970s, with the establishment of the journal *Cognitive Science* and the Cognitive Science Society, and the first published uses of the term "cognitive science" (e.g., Bobrow & Collins, 1975).

Cognitive science has thrived because the problem of understanding how the mind works is far too complex to be approached using ideas and methods from only one discipline. Many researchers whose primary backgrounds are in psychology, philosophy, neuroscience, linguistics, anthropology, and computer science have realized the advantages of tracking work in some of the other fields of cognitive science. Many successful projects have fruitfully combined methodologies from multiple fields, for example, research on inference that is both philosophical and computational, research on language that is both linguistic and neuroscientific, and research on culture that is both anthropological and psychological.

Naturally, cognitive science has also been used to investigate the mental processes required for the practice of science. The prehistory of the cognitive science of science goes back to philosophical investigation of scientific inference by Francis Bacon, David Hume, William Whewell, John Stuart Mill, and Charles Peirce. Modern cognitive science of science began only in the 1980s when various psychologists, philosophers, and computer scientists realized the advantages of taking a multidisciplinary approach to understanding scientific thinking. Pioneers include: Lindley Darden, Ronald Giere, and Nancy Nersessian in philosophy; Bruce Buchanan, Pat Langley, and Herbert Simon in computer modeling; and William Brewer, Susan Carey, Kevin Dunbar, David Klahr, and Ryan Tweney in experimental psychology. Extensive references are given in the next section. The earliest occurrence of the phrase "cognitive science of science" that I have been able to find is in Giere (1987), although the idea of applying cognitive psychology and computer modeling to scientific thinking goes back at least to Simon (1966).

This chapter provides a brief overview of what the component fields of cognitive science bring to the study of science, along with a sketch of the merits of combining methods. It also considers alternative approaches to science studies that are often antagonistic to the cognitive science of science, including formal philosophy of science and postmodernist history and sociology of science. I will argue that philosophy, history, and sociology of science can all benefit from ideas drawn from the cognitive sciences. Finally, I give an overview of the rest of the book by sketching how the cognitive science of science can investigate some of the most important aspects of the development of science, especially explanation, discovery, and conceptual change.

Approaches to the Cognitive Science of Science

It would take an encyclopedia to review all the different approaches to science studies that have been pursued. Much more narrowly and concisely, this section reviews what researchers from various fields have sought to contribute to the cognitive science of science.

My own original field is the philosophy of science, and I described in the preface how concern with the structure and growth of scientific knowledge led me to adopt ideas and methods from psychology and artificial intelligence, generating books and articles that looked at different aspects of scientific thinking (e.g., Thagard, 1988, 1992, 1999, 2000). Independently, other philosophers have looked to cognition to enhance understanding of science, including Lindley Darden (1983, 1991, 2006), David Gooding (1990), Ronald Giere (1988, 1999, 2010), and Nancy Nersessian (1984, 1992, 2008). Andersen, Barker, and Cheng (2006), Magnani (2001, 2009), and Shelley (2003) also combine philosophy of science, history of science, and cognitive psychology. Collections of work on philosophical approaches to the cognitive science of science include Giere (1992) and Carruthers, Stich, and Siegal (2002).

Philosophy of science is not just a beneficiary of cognitive science but also a major contributor to it. Since the 1600s work of Francis Bacon (1960), philosophers have investigated the nature of scientific reasoning and contributed valuable insights on such topics as explanation (Whewell 1967), causal reasoning (Mill 1970), and analogy (Hesse 1966). Philosophy of science was sidetracked during the logical positivist era by (1) a focus on formal logic as the canonical way of representing scientific information and (2) a narrow empiricism incapable of comprehending the theoretical successes of science. Logical positivism was as inimical to understanding scientific knowledge as behaviorism was to understanding thinking in general.

In response to logical positivism, Russell Hanson (1958), Thomas Kuhn (1962), and others spurred interest among philosophers in the history of science, but there was a dearth of tools richer than formal logic for examining science, although Hanson and Kuhn occasionally drew on insights from Gestalt psychology. In the 1980s, when philosophers looked to cognitive science for help in understanding historical developments, we brought to the cognitive science of science familiarity with many aspects

To complete this review of how the different fields of cognitive science contribute to the understanding of science, I need to include linguistics and anthropology. Unfortunately, I am not aware of much relevant research, although I can at least point to the work of Kertesz (2004) on the cognitive semantics of science, and to the work of Atran and Medin (2008) on folk concepts in biology across various cultures. Let me now return to why computer modeling is important for the cognitive science of science.

Methodology of Computational Modeling

What is the point of building computational models? One answer might come from the hypothetico-deductive view of scientific method, according to which science proceeds by generating hypotheses, deducing experimental predictions from them, and then performing experiments to see if the predicted observations occur. On this view, the main role of computational models is to facilitate deductions. There are undoubtedly fields such as mathematical physics and possibly economics where computer models play something like this hypothetico-deductive role, but their role in the cognitive sciences is much larger.

The hypothetico-deductive method is rarely applicable in biology, medicine, psychology, neuroscience, and the social sciences, where mathematically exact theories and precise predictions are rare. These sciences are better described by what I shall whimsically call the *mechanista* view of scientific method. Philosophers of science have described how many sciences aim for the discovery of mechanisms rather than laws, where a mechanism is a system of interacting parts that produce regular changes (e.g., Bechtel, 2008; Bechtel & Richardson, 1993; Bunge, 2003; Craver, 2007; Darden, 2006; Machamer, Darden, & Craver, 2000; Thagard, 2006a; Wimsatt, 2007). Biologists, for example, can rarely derive predictions from mathematically expressed theories, but they have been highly successful in describing mechanisms such as genetic variation and evolution by natural selection that have very broad explanatory scope. Similarly, I see cognitive science as primarily the search for mechanisms that can explain many kinds of mental phenomena such as perception, learning, problem solving, emotion, and language.

Computer modeling can be valuable for expressing, developing, and testing descriptions of mechanisms, at both psychological and neural levels

of explanation. In contemporary cognitive science, theories at the psychological level postulate various kinds of mental representations and processes that operate on them to generate thinking. For example, rule-based theories of problem solving, from Newell and Simon (1972) to Anderson (2007), postulate (1) representations of goals and if-then rules and (2) search processes involving selection and firing of rules. The representations are the parts and the processes are the interactions that together provide a mechanism that explains mental changes that accomplish tasks. Other cognitive science theories can also be understood as descriptions of mechanisms, for example, connectionist models that postulate simple neuronlike parts and processes of spreading activation that produce mental changes (Rumelhart & McClelland, 1986). Computational neuroscience now deals with much more biologically realistic neural entities and processes than connectionism, but the aim is the same: to describe the mechanisms that explain neuropsychological phenomena.

Expressing and developing such theoretical mechanisms benefits enormously from computational models. It is crucial to distinguish between theories, models, and programs. On the mechanista view, a theory is a description of mechanisms, and a model is a simplified description of the mechanisms postulated to be responsible for some phenomena. In computational models, the simplifications consist of proposing general kinds of data structures and algorithms that correspond to the parts and interactions that the theory postulates. A computer program produces a still more specific and idealized account of the postulated parts and interactions using data structures and algorithms in a particular programming language. For example, the theory of problem solving as rule application using means-ends reasoning gets a simplified description in a computational model with rules and goals as data structures and means-ends search as interactions. A computer program implements the model and theory in a particular programming language such as LISP or JAVA that makes it possible to run simulations. Theoretical neuroscience uses mathematically sophisticated programming tools such as MATLAB to implement computational models of neural structures and processes that approximate to mechanisms that are hypothesized to operate in brains.

Rarely, however, do computer modelers proceed simply from theory to model to program in the way just suggested. Rather, thinking about how to write a computer program in a familiar programming language

enables a cognitive scientist to express and develop ideas about what parts and interactions might be responsible for some psychological phenomena. Hence the development of cognitive theories, models, and programs is a highly interactive process in which theories stimulate the production of programs and vice versa. It is a mistake, however, to identify theories with programs, because any specific program will have many details arising from the peculiarities of the programming language used. Nevertheless, writing computer programs helps enormously to develop theoretical ideas expressed as computer models. The computer model provides a general analogue of the mechanisms postulated by the theory, and the program provides a specific, concrete, analogical instantiation of those mechanisms.

In the biological, social, and cognitive sciences, descriptions of mechanism are rarely so mathematical that predictions can be deduced, but running computer programs provides a looser way of evaluating theories and models. A computer program that instantiates a model that simplifies a theory can be run to produce simulations whose performance can be compared to actual behaviors, as shown in systematic observations, controlled behavioral experiments, or neurological experiments.

There are three degrees of evaluation that can be applied, answering the following questions about the phenomena to be explained:

1. Is the program capable of performing tasks like those that people have been observed doing?
2. Does the behavior of the program qualitatively fit with how people behave in experiments?
3. Does the behavior of the program quantitatively fit numerical data acquired in experiments?

Ideally, a computer program will satisfy all three of these tests, but often computer modeling is part of a theoretical enterprise that is well out in front of experimentation. In such cases, the program (and the model and theory it instantiates) can be used to suggest new experiments whose resulting data can be compared against the computer simulations. In turn, data that are hard to explain given currently available mechanisms may suggest new mechanisms that can be simulated by computer programs whose behaviors can once again be compared to those of natural systems. The three questions listed above apply to models of psychological

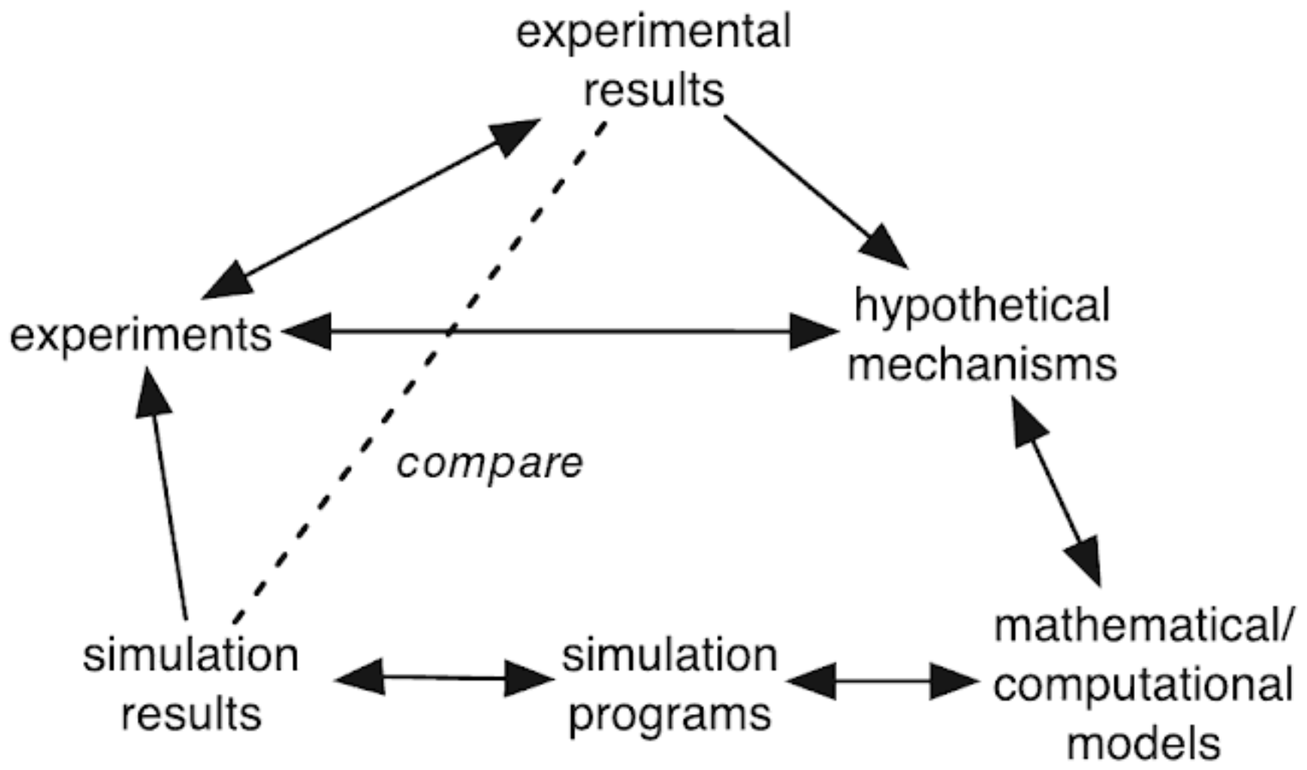

Figure 1.1
The role of computer models in developing and testing theories about mechanisms. Lines with arrows indicate causal influences in scientific thinking. The dashed line indicates the comparison between the results of experiments and the results of simulations.

behavior, but analogous questions can be asked about computational simulations of neural data.

The general interactive process of mechanism-based theory development using computational models is shown in figure 1.1, which portrays an interactive process with no particular starting point. Note that the arrows between mechanisms and models, and between models and simulations, are symmetrical, indicating that models can suggest mechanisms and programs can suggest models, as well as vice versa. In one typical pattern, experimental results prompt the search for explanatory mechanisms that can be specified using mathematical–computational models that are then implemented in computer programs. Simulations using these programs generate results that can be compared with experimental results. This comparison, along with insights gained during the whole process of generating mechanisms, models, and simulations, can in turn lead to ideas for new experiments that produce new experimental results.

Unified Cognitive Science Research

I have described philosophical, psychological, computational, and neuroscientific contributions to the understanding of science, but cognitive science at its best combines insights from all of its fields. We can imagine

what an ideal research project in the cognitive science of science would be like, one beyond the scope of any single researcher except perhaps Herbert Simon. Consider a team of researchers operating with a core set of theoretical ideas and multiple methodologies. Let "ASPECT" stand for some aspect of scientific thinking that has been little investigated. We can imagine a joint enterprise in which philosophers analyze historical cases of ASPECT, psychologists perform behavioral experiments on how adults and children do ASPECT, neuroscientists perform brain scans of people doing ASPECT, and computational modelers write programs that can simulate ASPECT. Linguists and anthropologists might also get involved by studying whether ASPECT varies across cultures. Representatives of all six fields could work together to generate and test theories about the mental structures and processes that enable people to accomplish ASPECT. My own investigations into the cognitive science of science do not have anything like the scope of this imaginary investigation of ASPECT, but they variously combine different parts of the philosophical, historical, psychological, computational, and neuroscientific investigation of scientific thinking.

Unified investigations in the cognitive science of science can be normative as well as descriptive. It is sometimes said that philosophy is normative, concerned with how things ought to be, in contrast to the sciences which are descriptive, concerned with how things are. This division is far too simple, because there are many applied sciences, from engineering to medicine to clinical and educational psychology, that aim to improve the world, not just to describe it (Hardy-Vallée & Thagard, 2008). Conversely, if the norms that philosophy seeks to develop are to be at all relevant to actual human practices, they need to be tied to descriptions of how the world, including the mind, generally works. I have elsewhere defended the naturalistic view that philosophy is continuous with science, differing in having a greater degree of generality and normativity (Thagard, 2009, 2010a). This book assumes the priority of scientific evidence and reasoning over alternative ways of fixing belief such as religious faith and philosophical thought experiments, but I argue for that assumption in Thagard (2010a, ch. 2).

The cognitive science of science can take from its philosophical component and also from its applied components a concern to be normative as well as descriptive. An interdisciplinary approach to science can aim not only to describe how science works, but also to develop norms for how it

alternative movement arose that managed to take over science studies programs at many universities. Sociologists of science produced a research program called the Sociology of Scientific Knowledge that abandoned the normative assessment of science in favor of purely sociological explanations of how science develops (e.g., Barnes, Bloor & Henry, 1996). Latour and Woolgar (1986) even called for a ten-year moratorium on cognitive explanations of science until sociologists had had a chance to explain all aspects of scientific development. That moratorium has long expired, and sociologists have obviously left lots of science to be explained. Moreover, some prominent proponents of postmodern sociology of science have made the shocking discovery that science and technology might even have *something* to do with reality (Latour, 2004).

In contrast to the imperialism of sociologists who think they can explain everything about scientific development, the cognitive science of science is friendly to sociological explanations. Power relations are undoubtedly an important part of scientific life, from the local level of laboratory politics to the national level of funding decisions. Like some analytic philosophers, some sociologists suffer from psychophobia, the fear of psychology, but cognitive approaches to science are compatible with the recognition of important social dimensions of science. For example, in my study of the development and acceptance of the bacterial theory of ulcers, I took into account social factors such as collaboration and consensus as well as psychological processes of discovery and evaluation (Thagard, 1999). Other works in the cognitive science of science have similarly attended to social dimensions (e.g., Dunbar, 1997; Giere, 1988). The cognitive and the social sciences should be seen as complements, not competitors, in a unified enterprise that might be called *cognitive social science*. Anthropology, sociology, politics, and economics can all be understood as requiring the integration of psychological and social mechanisms, as well as neural and molecular ones (Thagard, 2010d, forthcoming-c). Novel kinds of computer models are needed to explore how the behavior of groups can depend recursively on the behavior of individuals who think of themselves as members of groups. Agent-based models of social phenomena are being developed, but they are only just beginning to incorporate psychologically realistic agents (Sun, 2008a; Thagard, 2000, ch. 7, presents a cognitive-social model of scientific consensus). The aim of these models is not to reduce the social to the psychological and neural, but rather to show rich

interconnections among multiple levels of explanation. My hope is that future work on cognitive-social interactions will provide ways of simulating social aspects of science using techniques under development (Thagard, forthcoming-c).

Studies in the Cognitive Science of Science

In the rest of this book, however, I largely neglect social factors in science in order to concentrate on its philosophical, psychological, computational, and neural aspects. Even within the cognitive realm, the investigations reported here are selective, dealing primarily with explanation, discovery, and conceptual change. I understand science broadly to include medicine and technology, which are discussed in several of the chapters.

Part II considers cognitive aspects of explanation and related scientific practices concerned with the nature of theories and theory choice. After a brief overview that makes connections to related work, four chapters develop cognitive perspectives on the nature of explanation, mental models, theory choice, and resistance to scientific change. Climate change provides a case study where normative models of theory acceptance based on explanatory coherence are ignored because of psychological factors. This part also includes the most philosophical chapter in the book, arguing that coherence in science sometimes leads to truth.

Part III concerns scientific discovery understood as a psychological and neural process. Formal philosophy of science and sociological approaches have had little to say about how discoveries are made. In contrast, this part contains a series of studies about the psychological and neural processes that lead to breakthroughs in science, medicine, and technology.

Part IV shows how discoveries of new theories and explanations lead to conceptual change, ranging from the mundane addition of new concepts to the dramatic reorganizations required by scientific revolutions. Four chapters describe conceptual change in the fields of biology, psychology, and medicine.

Finally, Part V presents two new essays concerned with the nature of values and with the neural underpinnings of scientific thinking. The chapter on values shows how the cognitive science of science can integrate descriptive questions about how science works with normative questions about how it ought to work. The final chapter builds on Chris Eliasmith's

recent theory of semantic pointers to provide a novel account of the nature of scientific concepts such as *force*, *water*, and *cell*. (Please note that this book uses the following conventions: items in italics stand for concepts and items in quotes stand for words. For example, the concept *car* is expressed in English by the word "car" and applies to many things in the world, including my 2009 Honda Civic.)

The cognitive science of science inherits from the philosophy of science the problem of characterizing the structure and growth of scientific knowledge. It greatly expands the philosophical repertoire for describing the structure of knowledge by introducing much richer and empirically supported accounts of the nature of concepts, rules, mental models, and other kinds of representations. Even greater is the expansion of the repertoire of mechanisms for explaining the growth of scientific knowledge, through computationally rich and experimentally testable models of the nature of explanation, coherence, theory acceptance, inferential bias, concept formation, hypothesis discovery, and conceptual change. Adding an understanding of the psychological and neural processes that help to generate and establish scientific knowledge does not undercut philosophical concerns about normativity and truth, nor need it ignore the social processes that are also important for the development of scientific knowledge. Although the cognitive science of science is only a few decades old, I hope that the essays in this book, along with allied work by others, show its potential for explaining science.

II Explanation and Justification

Second, explanation is an important part of science education. Teachers need to convey to students how science provides explanations and give them a taste of how scientific explanations work. Cognitive science should be able to integrate insights from philosophy, psychology, neuroscience, and computer modeling to elucidate what needs to go on in the minds of students as they learn at least to appreciate scientific explanations and at best to be able to develop new ones of their own.

Third, explanation is not only an intrinsic goal of science, but is intimately connected with another important goal—truth. Nowadays, people, especially social scientists, are sometimes embarrassed by the suggestion that science can achieve truth, interpreting talk of reality as a vestige of naive philosophical ideas that expired with Kant or with twentieth-century postmodernism. At the other extreme, some philosophers assume that science is primarily aimed just at truth, with explanation at best a sideshow. I reject both these views, and the chapters in Part II present a picture in which explanation is a key aspect of justifying the acceptance of hypotheses and theories. If theory choice is governed by inference to the best explanation, as chapter 5 assumes, then explanation is directly relevant to the question of what theories we should accept as true. Whether such theories really are true is a matter for discussion, as it is clear that many theories have been accepted by scientific communities that were later found to be defective. Chapter 6 provides a stronger connection between explanation, justification, and truth.

Of the fields of cognitive science, philosophy has the most ancient concern with the nature of explanation, going back at least to Aristotle. The philosophy of science took off in the 1800s with incisive discussions of explanation and explanatory reasoning by William Whewell, John Stuart Mill, and Charles Peirce. Peirce (1992) coined the term "abduction" for a kind of inference that generates and/or evaluates explanatory hypotheses. In the middle of the 1900s, the logical positivists developed a theory of explanation as deduction from laws that is still influential in philosophical circles (Hempel, 1965). More recently, many philosophers concerned with explanation in biology and cognitive science have highlighted the role that descriptions of mechanisms play in scientific explanations, as I reviewed in discussing the mechanista approach in chapter 1; see also Bechtel and Abrahamsen (2005). Woodward (2009) gives a good,

brief overview of philosophical work on explanation. Kitcher and Salmon (1989) provide a useful older collection on the philosophy of scientific explanation.

Explanation has been an important concern for cognitive, developmental, and social psychologists. Cognitive psychologists have seen the relevance of explanation to the general theory of concepts, with some arguing that the functions of concepts include not just classification of objects but also explanation of why things happen (e.g., Lombrozo, 2009; Medin, 1989; Murphy, 2002). For example, saying that something is a bear can explain why it eats fish. Developmental psychologists have interpreted concept acquisition in children as partly aimed at providing explanations of why things happen (Gopnik, 1998; Keil, 2006). Social psychologists have long been concerned with how people explain the actions of others, a process they call attribution (Kelley, 1973). All of these investigations are relevant to scientific explanation, assuming some commonality between it and explanation in everyday life. See Keil and Wilson (2000) for a collection on the psychology of explanation.

Explanation was an important topic in artificial intelligence in the 1980s and 1990s (e.g., Minton et al., 1989), but it seems to have declined in importance as researchers moved to more statistical approaches. This decline is unfortunate, because AI needs to replicate the most sophisticated kinds of thinking, including what scientists do when they explain things. As chapter 3 describes, much work in AI has been limited to a deductive view of explanation, which at best captures only some kinds of scientific explanation.

Anthropologists and linguists have not, to my knowledge, done much to investigate the nature of scientific explanation. Nor has experimental and theoretical neuroscience yet said much about explanation, but some of the chapters below begin to fill this gap.

For Part II, I have chosen four recent papers as contributions to the cognitive science of explanation. Chapter 3 provides an overview of computational models of explanation, reviewing ones based on deduction, schemas, analogy, probability, and neural networks. It presents a model of how a simple form of abductive inference can be performed in a biologically plausible neural network. Chapter 4 shows how the same kind of neural network approach is relevant to explaining how high-level

cognitive processes involving mental models can be understood neuro-computationally. This chapter also addresses the question of the extent to which cognition is embodied.

The next two chapters are concerned with philosophical questions about justification and truth, but they approach these from the cognitive perspective that explanation and inference are mental processes. Chapter 5 uses computational models of both correct and biased explanatory coherence to explain the nature of current debates about climate change. Ideally, claims about whether there is global warming and whether it is the result of human activities should be based solely on whether the hypotheses in question explain the evidence. These issues, however, are fraught with economic and political problems, so it is not surprising that people's thinking can be biased by their motivations. Cognitive science can explain not only how people think when they are doing it right, but also how thinking is often distorted by goals extraneous to the scientific aims of explanation and truth. Later, chapter 14 provides a similar account of resistance to Darwin's theory of evolution.

Finally, chapter 6, the most philosophical one in this book, argues that, under certain conditions, it is legitimate to conclude that theories that provide the best available evidence do indeed approximate the truth. This chapter provides a justification for the philosophical position called scientific realism, according to which science aims and sometimes succeeds in achieving truth. Thagard (2010a) provides a more general defense of realism.

3 Models of Scientific Explanation

Paul Thagard and Abninder Litt

Explanation

Explanation of why things happen is one of humans' most important cognitive operations. In everyday life, people are continually generating explanations of why other people behave the way they do, of why they get sick, of why computers or cars are not working properly, and of many other puzzling occurrences. More systematically, scientists develop theories to provide general explanations of physical phenomena such as why objects fall to Earth, chemical phenomena such as why elements combine, biological phenomena such as why species evolve, medical phenomena such as why organisms develop diseases, and psychological phenomena such as why people sometimes make mental errors.

This chapter reviews computational models of the cognitive processes that underlie these kinds of explanations of *why* events happen. It is not concerned with another sense of explanation that just means clarification, as when someone explains the U.S. Constitution. The focus will be on scientific explanations, but more mundane examples will occasionally be used, on the grounds that the cognitive processes for explaining why events happen are much the same in everyday life and in science, although scientific explanations tend to be more systematic and rigorous than everyday ones. In addition to providing a concise review of previous computational models of explanation, this chapter describes a new neural network model that shows how explanations can be performed by multimodal distributed representations.

Before proceeding with accounts of particular computational models of explanation, let us characterize more generally the three major processes involved in explanation and the four major theoretical approaches that

have been taken in computational models of it. The three major processes are: providing an explanation from available information, generating new hypotheses that provide explanations, and evaluating competing explanations. The four major theoretical approaches are: deductive, using logic or rule-based systems; schematic, using explanation patterns or analogies; probabilistic, using Bayesian networks; and neural, using networks of artificial neurons. For each of these theoretical approaches, it is possible to characterize the different ways in which the provision, generation, and evaluation of explanations are understood computationally.

The processes of providing, generating, and evaluating explanations can be illustrated with a simple medical example. Suppose you arrive at your doctor's office with a high fever, headache, extreme fatigue, a bad cough, and major muscle aches. Your doctor will probably tell you that you have been infected by the influenza virus, with an explanation like:

People infected by the flu virus often have the symptoms you describe.
You have been exposed to and infected by the flu virus.
So, you have these symptoms.

If influenza is widespread in your community and your doctor has been seeing many patients with similar symptoms, it will not require much reasoning to provide this explanation by stating the flu virus as the likely cause of your symptoms.

Sometimes, however, a larger inferential leap is required to provide an explanation. If your symptoms also include a stiff neck and confusion, your doctor may make the less common and more serious diagnosis of meningitis. This diagnosis requires generating the hypothesis that you have been exposed to bacteria or viruses that have infected the lining surrounding the brain. In this case, the doctor is not simply applying knowledge already available to provide an explanation, but generating a hypothesis about you that makes it possible to provide an explanation. This hypothesis presupposes a history of medical research that led to the identification of meningitis as a disease caused by particular kinds of bacteria and viruses, research that required the generation of new general hypotheses that made explanation of particular cases of the disease possible.

In addition to providing and generating explanations, scientists and ordinary people sometimes need to evaluate competing explanations. If

Many computational models in artificial intelligence have presupposed that explanation is deductive, including ones found in logic programming, truth maintenance systems, explanation-based learning, qualitative reasoning, and in some approaches to abduction (a form of inference that involves the generation and evaluation of explanatory hypotheses). See, for example, Russell and Norvig (2003), Bylander et al. (1991), and Konolige (1992). These AI approaches are not intended as models of human cognition, but see Bringsjord (2008) for discussion of the use of formal logic in cognitive modeling.

Deductive explanation also operates in rule-based models, which have been proposed for many kinds of human thinking (Anderson, 1983, 1993, 2007; Holland et al., 1986; Newell & Simon, 1972; Newell, 1990). A rule-based system is a set of rules with an "if" part consisting of conditions (antecedents) and a "then" part consisting of actions (consequents). Rule-based systems have often been used to model human problem solving in which people need to figure out how to get from a starting state to a goal state by applying a series of rules. This is a kind of deduction, in that the application of rules in a series of if-then inferences amounts to a series of applications of the rule of deductive inference, modus ponens, which licenses inferences from *p* and *if p then q* to *q*. Most rule-based systems, however, do not always proceed from starting states to goal states, but can also work backward from a goal state to find a series of rules that can be used to get from the starting state to the goal state.

Explanation can be understood as a special kind of problem solving, in which the goal state is a target to be explained. Rule-based systems do not have the full logical complexity to express the laws required for Hempel's model of explanation, but they can perform a useful approximation. For instance, the medical example used in the introduction can be expressed by a rule like:

If *X* has influenza, then *X* has fever, cough, and aches.
Paul has influenza.

Paul has fever, cough, and aches.

Modus ponens provides the connection between the rule and what is to be explained. In more complex cases, the connection would come from a sequence of applications of modus ponens as multiple rules get applied. In contrast to Hempel's account in which an explanation is a static

argument, rule-based explanation is usually a dynamic process involving application of multiple rules. For a concrete example of a running program that accomplishes explanations in this way, see the PI ("processes of induction") model of Thagard (1988; code is available at http://cogsci.uwaterloo.ca). The main scientific example to which PI has been applied is the discovery of the wave theory of sound, which occurs in the context of an attempt to explain why sounds propagate and reflect.

Thus rule-based systems can model the provisions of explanations construed deductively, but what about the generation and evaluation of explanations? A simple form of abductive inference that generates hypotheses can be modeled as a kind of backward chaining. Forward chaining involves running rules forward in the deductive process that proceeds from the starting state toward a goal to be solved. Backward chaining occurs when a system works backward from a goal state to find rules that could produce it from the starting state. Human problem solving on tasks such as solving mathematics problems often involves a combination of forward and backward reasoning, in which a problem solver looks both at the how the problem is described and the answer that is required, attempting to make them meet. At the level of a single rule, backward chaining has the form: goal *G* is to be accomplished; there is the rule if *A* then *G*, that is, action *A* would accomplish *G*; so set *A* as a new subgoal to be accomplished. Analogously, people can backchain to find a possible explanation: fact *F* is to be explained; there is a rule if *H* then *F*, that is, hypothesis *H* would explain *F*; so hypothesize that *H* is true. Thus, if you know that Paul has fever, aches, and a cough, and you know the rule that if *X* has influenza, then *X* has fever, cough, and aches, then you can run the rule backward to produce the hypothesis that Paul has influenza.

The computational PI model performs this simple kind of hypothesis generation, but it also can generate other kinds of hypotheses (Thagard, 1988). For example, from the observation that the orbit of Uranus is perturbed, and the rule that if a planet has another planet near it then its orbit is perturbed, PI infers that there is some planet near Uranus; this is called "existential abduction." PI also performs abduction to rules that constitute the wave theory of sound: the attempt to explain why an arbitrary sound propagates generates not only the hypothesis that it consists of a wave but the general theory that all sounds are waves. PI also performs

a kind of analogical abduction, a topic discussed in the next section on schemas.

Abductive inference that generates explanatory hypotheses is an inherently risky form of reasoning because of the possibility of alternative explanations. Inferring that Paul has influenza because it explains his fever, aches, and cough is risky because other diseases such as meningitis can cause the same symptoms. People should only accept an explanatory hypothesis if it is better than its competitors, a form of inference that philosophers call "inference to the best explanation" (Harman, 1973; Lipton, 2004). The PI cognitive model performs this kind of inference by taking into account three criteria for the best explanation: consilience, which is a measure of how much a hypothesis explains; simplicity, which is a measure of how few additional assumptions a hypothesis needs to carry out an explanation; and analogy, which favors hypotheses whose explanations are analogous to accepted ones. A more psychologically elegant way of performing inference to the best explanation, the ECHO model, is described below in the section on neural networks. Neither the PI nor the ECHO way of evaluating competing explanations requires that explanations be deductive.

In artificial intelligence, the term "abduction" is often used to describe inference to the best explanation as well as the generation of hypotheses. In actual systems, these two processes can be continuous, for example in the PEIRCE tool for abductive inference described by Josephson and Josephson (1994, p. 95). This is primarily an engineering tool rather than a cognitive model, but we mention it here as another approach to generating and evaluating scientific explanations, in particular medical ones involving interpretation of blood tests. The PEIRCE system accomplishes the goal of generating the best explanatory hypothesis by achieving three subgoals:

1. generation of a set of plausible hypotheses;
2. construction of a compound explanation for all the findings; and
3. criticism and improvement of the compound explanation.

PEIRCE employs computationally effective algorithms for each of these subgoals, but it does not attempt to do so in a way that corresponds to how people accomplish them.

Schema and Analogy Models

In ordinary life, and in many areas of science less mathematical than physics, the relation between what is explained and what does the explaining is usually looser than deduction. An alternative conception of this relation is provided by understanding an explanation as the application of a causal schema, which is a pattern that describes the relation between causes and effects. For example, cognitive science uses a general explanation schema with the following structure (Thagard, 2005a):

Explanation target: Why do people have a particular kind of **intelligent behavior**?

Explanatory pattern:
People have mental **representations.**
People have algorithmic **processes** that operate on those **representations.**
The **processes**, applied to the **representations**, produce the **behavior**.

This schema provides explanations when the terms shown in boldface are filled in with specifics, and subsumes schemas that describe particular kinds of mental representations such as concepts, rules, and neural networks. Philosophers of science have discussed the importance of explanation schemas or patterns (Kitcher, 1993; Thagard, 1999).

A computational cognitive model of explanation schemas was developed in the SWALE project (Schank, 1986; Leake, 1992). This project modeled people's attempts to explain the unexpected 1984 death of a racehorse, Swale. Given an occurrence, the program SWALE attempts to fit it into memory. If a problem arises indicating an anomaly, then the program attempts to find an explanation pattern stored in memory. The explanation patterns are derived from previous cases, such as other unexpected deaths. If SWALE finds more than one relevant explanation pattern, it evaluates them to determine which is most relevant to the intellectual goals of the person seeking understanding. If the best-explanation pattern does not quite fit the case to be explained, it can be tweaked (adapted) to provide a better fit, and the tweaked version is stored in memory for future use. The explanation patterns in SWALE's database included both general schemas such as *exertion + heart defect causes fatal heart attack* and particular examples, which are used for case-based reasoning, a kind of analogical thinking. Leake (1992) describes how competing explanation patterns can

be evaluated according to various criteria, including a reasoner's pragmatic goals.

Explaining something by applying a general schema involves the same processes as explaining using analogies. In both cases, reasoning proceeds as follows:

Identify the case to be explained.

Search memory for a matching schema or case.

Adapt the found schema or case to provide an explanation of the case to be explained.

In deductive explanation, there is a tight logical relation between what is explained and the sentences that imply it, but in schematic or analogical explanation there need only be a roughly specified causal relation.

Falkenhainer (1990) describes a program, PHINEAS, that provides analogical explanations of scientific phenomena. The program uses Forbus's (1984) qualitative process theory to represent and reason about physical change, and is provided with knowledge about liquid flow. When presented with other phenomena to be explained such as osmosis and heat flow, it can generate new explanations analogically by computing similarities in relational structure, using the Structure Mapping Engine (Falkenhainer, Forbus & Gentner, 1989). PHINEAS operates in four stages: access, mapping/transfer, qualitative simulation, and revision. For example, it can generate an explanation of the behavior of a hot brick in cold water by analogy to what happens when liquid flows between two containers. Another computational model that generates analogical explanations is the PI system (Thagard, 1988), mentioned above, which simulates the discovery of the wave theory of sound by analogy to water waves.

Thus computational models of explanation that rely on matching schematic or analogical structures based on causal fit provide alternatives to models of deductive explanation. These two approaches are not competing theories of explanation, because explanation can take different forms in different areas of science. In areas such as physics that are rich in mathematically expressed knowledge, deductive explanations may be available. But in more qualitative areas of science and everyday life, explanations are usually less exact and may be better modeled by application of causal schemas or as a kind of analogical inference.

of feedback loops. For example, marriage breakdown often occurs because of escalating negative affect, in which the negative emotions of one partner produce behaviors that increase the negative emotions of the other, which then produce behavior that increases the negative emotions of the first partner (Gottman et al., 2003). Such feedback loops are also common in biochemical pathways needed to explain disease (Thagard, 2003). Fourth, probability by itself is not adequate to capture people's understanding of causality, as argued in the last section of this chapter. Hence it is not at all obvious that Bayesian networks are the best way to model explanation by human scientists. Even in statistically rich fields such as the social sciences, scientists rely on an intuitive, nonprobabilistic sense of causality of the sort discussed below.

Neural Network Models

The most important approach to cognitive modeling not yet discussed here employs artificial neural networks. Applying this approach to high-level reasoning faces many challenges, particularly in representing the complex kinds of information contained in scientific hypotheses and causal relations. Thagard (1989) provided a neural network model of how competing scientific explanations can be evaluated, but did so using a localist network in which entire propositions were represented by single artificial neurons and in which relations between propositions are represented by excitatory and inhibitory links between the neurons. Although this model provides an extensive account of explanation evaluation, which is reviewed below, it reveals nothing about what an explanation is or how explanations are generated. Neural network modelers have been concerned mostly with applications to low-level psychological phenomena such as perception, categorization, and memory, rather than high-level ones such as problem solving and inference (O'Reilly & Munakata, 2000). However, this section shows how we can construct a neurologically complex model of explanation and abductive inference.

One benefit of attempting neural analyses of explanation is that it becomes possible to incorporate multimodal aspects of cognitive processing that tend to be ignored by the deductive, schematic, and probabilistic perspectives. Thagard (2007a) describes how both explainers and explanation targets are sometimes represented nonverbally. In medicine, for

Figure 3.2
The process of abductive inference. (From Thagard, 2007a.)

example, doctors and researchers may employ visual hypotheses (say, about the shape and location of a tumor) to explain observations that can be represented using sight, touch, and smell as well as words. Moreover, the process of abductive inference has emotional inputs and outputs, because it is usually initiated when an observation is found to be surprising or puzzling, and it often results in a sense of pleasure or satisfaction when a satisfactory hypothesis is used to generate an explanation. Figure 3.2 provides an outline of this process. Let us now look at an implementation of a neural network model of this sketch.

The model of abduction described here follows the Neural Engineering Framework (NEF) outlined by Eliasmith and Anderson (2003), and is implemented using the MATLAB-based NEF simulation software *NESim*. The NEF proposes three basic principles of neural computation (Eliasmith & Anderson, 2003, p. 15):

1. Neural representations are defined by a combination of nonlinear encoding and linear decoding.
2. Transformations of neural representations are linearly decoded functions of variables that are represented by a neural population.
3. Neural dynamics are characterized by considering neural representations as control theoretic state variables.

These principles are applied to a particular neural system by identifying the interconnectivity of its subsystems, neuron response functions, neuron tuning curves, subsystem functional relations, and overall system behavior. For cognitive modeling, the NEF is useful because it provides a mathematically rigorous way of building more realistic neural models of cognitive functions.

The NEF characterizes neural populations and activities in terms of mathematical representations and transformations. The complexity of a representation is constrained by the *dimensionality* of the neural population that represents it. In rough terms, a single dimension in such a representation can correspond to one discrete "aspect" of that representation (e.g., speed and direction are the dimensional components of the vector quantity velocity). A hierarchy of representational complexity thus follows from neural activity defined in terms of one-dimensional scalars; vectors, with a finite but arbitrarily large number of dimensions; or functions, which are essentially *continuous* indexings of vector elements, thus ranging over infinite dimensional spaces.

The NEF provides for arbitrary computations to be performed in biologically realistic neural populations and has been successfully applied to phenomena as diverse as lamprey locomotion (Eliasmith & Anderson, 2003), path integration by rats (Conklin & Eliasmith, 2005), and the Wason card selection task (Eliasmith, 2005a). The Wason task model, in particular, is structured very similarly to the model of abductive inference discussed here. Both employ *holographic reduced representations* (HRRs), a high-dimensional form of distributed representation.

First developed by Plate (2003), HRRs combine the neurological plausibility of distributed representations with the ability to maintain complex, embedded structural relations in a computationally efficient manner. This ability is common in symbolic models and is often singled out as deficient in distributed connectionist frameworks; for a comprehensive review of HRRs in the context of the distributed versus symbolic representation debate, see Eliasmith and Thagard (2001). HRRs consist of high-dimensional vectors combined via multiplicative operations, and are similar to the tensor products used by Smolensky (1990) as the basis for a connectionist model of cognition. But HRRs have the important advantage of *fixed dimensionality*: the combination of two n-dimensional HRRs produces another n-dimensional HRR, rather than the $2n$ or even n^2 dimensionality one would obtain using tensor products. This avoids the explosive computational resource requirements of tensor products to represent arbitrary, complex structural relationships.

HRR representations are constructed through the multiplicative *circular convolution* (denoted by $\otimes$) and are decoded by the approximate inverse operation, *circular correlation* (denoted by #). The details of these operations

are given in the appendixes of Eliasmith and Thagard (2001), but in general if $C = A \otimes B$ is encoded, then $C\#A \approx B$ and $C\#B \approx A$. The approximate nature of the unbinding process introduces a degree of noise, proportional to the complexity of the HRR encoding in question and in inverse proportion to the dimensionality of the HRR (Plate, 2003). As noise tolerance is a requirement of any neurologically plausible model, this loss of representation information is acceptable, and the "cleanup" method of recognizing encoded HRR vectors using the dot product can be used to find the vector that best fits what was decoded (Eliasmith & Thagard, 2001). Note that HRRs may also be combined by simple superposition (i.e., addition): $P = Q \otimes R + X \otimes Y$, where $P\#R \approx Q$, $P\#X \approx Y$, and so on. The operations required for convolution and correlation can be implemented in a recurrent connectionist network (Plate, 2003) and in particular under the NEF (Eliasmith, 2005a).

In brief, the new model of abductive inference involves several large, high-dimensional populations to represent the data stored via HRRs and learned HRR transformations (the main output of the model), and a smaller population representing emotional valence information (abduction only requires considering emotion scaling from surprise to satisfaction, and hence needs only a single dimension represented by as few as 100 neurons to represent emotional changes). The model is initialized with a base set of causal encodings consisting of 100-dimensional HRRs combined in the form

$$antecedent \otimes a + relation \otimes causes + consequent \otimes b,$$

as well as HRRs that represent the successful explanation of a target x ($expl \otimes x$). For the purposes of this model, only six different "filler" values were used, representing three such causal rules (a causes b, c causes d, and e causes f). The populations used have between 2,000 and 3,200 neurons each and are 100- or 200-dimensional, which is at the lower end of what is required for accurate HRR cleanup (Plate, 2003). More rules and filler values would require larger and higher-dimensional neural populations, an expansion that is unnecessary for a simple demonstration of abduction using biologically plausible neurons.

Following detection of a surprising b, which could be an event, proposition, or any sensory or cognitive data that can be represented via neurons, the change in emotional valence spurs activity in the output population

toward generating a hypothesized explanation. This process involves employing several neural populations (representing the memorized rules and HRR convolution/correlation operations) to find an antecedent involved in a causal relationship that has *b* as the consequent. In terms of HRRs, this means producing (*rule # antecedent*) for ([*rule # relation* ≈ *causes*] and [*rule # consequent* ≈ *b*]). This production is accomplished in the 2,000-neuron, 100-dimensional output population by means of associative learning through recurrent connectivity and connection-weight updating (Eliasmith, 2005). As activity in this population settles, an HRR cleanup operation is performed to obtain the result of the learned transformation. Specifically, some answer is "chosen" if the cleanup result matches one encoded value significantly more than any of the others (i.e., is above some reasonable threshold value).

After the successful generation of an explanatory hypothesis, the emotional valence signal is reversed from surprise (which drove the search for an explanation) to what can be considered pleasure or satisfaction derived from having arrived at a plausible explanation. This in turn induces the output population to produce a representation corresponding to the successful dispatch of the explanandum *b*: namely, the HRR $expl_b = expl \otimes b$. Upon settling, it can thus be said that the model has accepted the hypothesized cause obtained in the previous stage as a valid explanation for the target *b*. Settling completes the abductive inference: emotional valence returns to a neutral level, which suspends learning in the output population and causes population firing to return to basal levels of activity.

Figure 3.3 shows the result of performing the process of abductive inference in the neural model, with activity in the output population changing with respect to changing emotional valence, and vice versa. The output population activity is displayed by dimension, rather than individual neuron, since the 100-dimensional HRR output of the neural ensemble as a whole is the real characterization of what is being represented. The boxed sets of numbers represent the results of HRR cleanups on the output population at different points in time; if one value reasonably dominates over the next few largest, it can be taken to be the "true" HRR represented by the population at that moment. In the first stage, the high emotional valence leads to the search for an antecedent of a causal rule" for *b*, the surprising explanandum. The result is an HRR cleanup best fitting to *a*, which is indeed the correct response. Reaching

accepted, whereas if a unit ends up with negative activation, the proposition it represents is rejected.

ECHO has been used to model numerous cases in the history of science, and has also inspired experimental research in social and educational psychology (Read & Marcus-Newhall, 1993; Schank & Ranney, 1991). The model shows how a very high-level kind of cognition, evaluating complex theories, can be performed by a simple neural network performing parallel constraint satisfaction. ECHO has a degree of psychological plausibility, but for neurological plausibility it pales in comparison to the NEF model of abduction described earlier in this section. The largest ECHO model uses only around 200 units to encode the same number of propositions, whereas the NEF model uses thousands of spiking neurons to encode a few causal relations. Computationally, this seems inefficient, but of course the brain has many billions of neurons that provide its distributed representations.

How might one implement comparative theory evaluation as performed by ECHO within the NEF framework? Thagard and Aubie (2008) use the NEF to encode ECHO networks by generating a population of thousands of neurons. Parallel constraint satisfaction is performed by transformations of neurons that carry out approximately the same calculations that occur more directly in ECHO's localist neural networks. Hence it is now possible to model the evaluation of competing explanations using more biologically realistic neural networks.

Causality

Like most other models of explanation, these neural network models presuppose some understanding of causality. In one sense that is common in both science and everyday life, to explain something involves stating its cause. For example, when people have influenza, the virus that infects them is the cause of their symptoms such as fever. But what is a cause? Philosophical theories of explanation correlate with competing theories of causality; for example, the deductive view of explanation fits well with the Humean understanding of causality as constant conjunction. If all *A*s are *B*s, then someone can understand how being an *A* can cause and explain being a *B*. Unfortunately, universality is not a requisite of either explanation or causality. Smoking causes lung cancer, even though many smokers

never get lung cancer, and some people with lung cancer have never smoked. Schematic models of explanation presuppose a primitive concept of causation without being able to say much about it. Probability theory may look like a promising approach to causality in that causes make their effects more probable than they would be otherwise, but such increased probability may be accidental or the result of some common cause. For example, the probability of someone drowning is greater on a day when much ice cream is consumed, but that is because of the common cause that more people go swimming and more people eat ice cream on hot days. Sorting out causal probabilistic information from misleading correlations requires much information about probability and independence that people usually lack.

Thagard (2007a) conjectured that it might be possible to give a neural network account of how organisms understand causality. Suppose, in keeping with research on infants' grasp of causality, that cause is a preverbal concept based on perception and motor control (Baillargeon, Kotovsky & Needham, 1995; Mandler, 2004). Consider an infant a few months old, lying on its back and swiping at a mobile suspended over its head. The infant has already acquired an image schema of the following form:

perception of situation + motor behavior ⇒ perception of new situation.

Perhaps this schema is innate, but alternatively it may have been acquired from very early perceptual–motor experiences in which the infant acted on the world and perceived the resultant changes. A simple instance of the schema would be:

stationary object + hand hitting object ⇒ moving object.

The idea of a preverbal image schema for causality is consistent with the views of some philosophers that manipulability and intervention are central features of causality (Woodward, 2004). The difference between *A* causing *B* and *A* merely being correlated with *B* is that manipulating *A* also manipulates *B* in the former case but not the latter. Conceptually, the concepts of manipulation and intervention seem to presuppose the concept of causation, because making something happen is on the surface no different from causing it to happen. However, although there is circularity at the verbal level, psychologically it is possible to break out of the circle by supposing that people have from infancy a neural encoding of the causality

image schema described above. This nonverbal schema is the basis for understanding the difference between one event making another event happen and one event just occurring after the other.

The causality image schema is naturally implemented within the Neural Engineering Framework used to construct the model of abductive inference. Neural populations are capable of encoding both perceptions and motor behaviors, and are also capable of encoding relations between them. In the model of abductive inference described in the last section, *cause* (*c*, *e*) was represented by a neural population that encodes an HRR vector that captures the relation between a vector representing *c* and a vector representing *e*, where both of these can easily be nonverbal perceptions and actions as well as verbal representations. In the NEF model of abduction, there is no real understanding of causality, because the vector was generated automatically. In contrast, it is reasonable to conjecture that people have neural populations that encode the notion of causal connection as the result of their very early preverbal experience with manipulating objects. Because the connection is based on visual and kinesthetic experiences, it cannot be adequately formulated linguistically, but it provides the intellectual basis for the more verbal and mathematical characterizations of causality that develop later.

If this account of causality is correct, then a full cognitive model of explanation cannot be purely verbal or probabilistic. Many philosophers and cognitive scientists currently maintain that scientific explanation of phenomena consists in providing mechanisms that produce those phenomena (e.g., Bechtel & Abrahamsen, 2005; Sun, Coward & Zenzen, 2005). A mechanism is a system of objects whose interactions regularly produce changes. All of the computational models described in this chapter are mechanistic, although they differ in what they take to be the parts and interactions that are central to explaining human thinking; for the neural network approaches, the computational mechanisms are also biological ones. But an understanding of mechanism presupposes an understanding of causality, in that there must be a relation between the interactions of the parts that constitutes production of the relevant phenomena. Because scientific explanation depends on the notion of causality, and because understanding of causality is in part visual and kinesthetic, future comprehensive cognitive models of explanation will need to incorporate neural network simulations of people's nonverbal understanding of causality.

Table 3.1
Summary of approaches to computational modeling of explanation.

	Target of explanation	Explainers	Relation between target and explainers	Mode of generation
Deductive	sentence	sentences	deduction	backward chaining
Schema	sentence	pattern of sentences	fit	search for fit, schema generation
Probabilistic	variable node	Bayesian network	conditional probability	Bayesian learning
Neural network	neural group: multimodal representation	neural groups	gated activation, connectivity	search, associative learning

Conclusion

This chapter has reviewed four major computational approaches to understanding scientific explanations: deductive, schematic, probabilistic, and neural network. Table 3.1 summarizes the different approaches to providing and generating explanations. To some extent, the approaches are complementary rather than competitive, because explanation can take different forms in different areas of science and everyday life. However, at the root of scientific and everyday explanation is an understanding of causality represented nonverbally in human brains by populations of neurons encoding how physical manipulations produce sensory changes. Another advantage of taking a neural network approach to explanation is that it becomes possible to model how abductive inference, the generation of explanatory hypotheses, is a process that is multimodal, involving not only verbal representations but also visual and emotional ones that constitute inputs and outputs to reasoning.

4 How Brains Make Mental Models

Introduction

Mental models are psychological representations that have the same relational structure as what they represent. They have been invoked to explain many important aspects of human reasoning, including deduction, induction, problem solving, language understanding, and human–machine interaction. But the nature of mental models and of the processes that operate on them has not always been clear from the psychological discussions. The main aim of this chapter is to provide a neural account of mental models by describing some of the brain mechanisms that produce them.

The neural representations required to understand mental models are also valuable for providing new understanding of how minds perform abduction, a kind of inference that generates and/or evaluates explanatory hypotheses. Considering the neural mechanisms that support abductive inference makes it possible to address several aspects of abduction, some first proposed by Charles Peirce, that have largely been neglected in subsequent research. These aspects include the generation of new ideas, the role of emotions such as surprise, the use of multimodal representations to produce "embodied abduction," and the nature of the causal relations that are required for explanations.

The suggestion that abductive inference is embodied raises issues that have been very controversial in recent discussions in psychology, philosophy, and artificial intelligence. This chapter argues that the role of emotions and multimodal representations in abduction supports a moderate thesis about the role of embodiment in human thinking, but not an extreme thesis that proposes embodied action as an alternative to the computational-representational understanding of mind.

relation in the sentential abductive schema. Presumably it must be more than material implication, but what more is required? Logic-based approaches to abduction tend to assume that explanation is a matter of deduction, but philosophical discussions show that deduction is neither necessary nor sufficient for explanation (e.g., Salmon, 1989). I think that good explanations exploit causal mechanisms, but what constitutes the causal relation between what is explained and what gets explained? I aim to show that all of these difficult aspects of abduction—the role of surprise and insight, the generation of new ideas, and the nature of causality—can be illuminated by consideration of neural mechanisms.

Terminological note: Magnani (2009) writes of "non-explanatory abduction," which strikes me as self-contradictory. Perhaps there is a need for a new term describing a kind of generalization of abduction to cover other kinds of backward or inverse reasoning such as generating axioms from desired theorems, but let me propose to call this generalized abduction "gabduction" and retain the term "abduction" for Peirce's idea of the generation and evaluation of explanatory hypotheses.

Neural Representation and Processing

A full and rigorous description of current understanding of the nature of neural representation and processing is beyond the scope of this chapter, but I will provide an introductory sketch (for fuller accounts, see such sources as Churchland & Sejnowski, 1992; Dayan & Abbott, 2001; Eliasmith & Anderson, 2003; O'Reilly & Munakata, 2000; and Thagard, 2010a).

The human brain contains around 100,000,000,000 neurons, each of which has many thousands of connections with other neurons. These connections are either excitatory (the firing of one neuron increases the firing of the one it is connected to) or inhibitory (the firing of one neuron decreases the firing of the one it is connected to). A collection of neurons that are richly interconnected is called a neural population (or group, or ensemble). A neuron fires when it has accumulated sufficient voltage as the result of the firing of the neurons that have excitatory connections to it. Typical neurons fire around 100 times per second, making them vastly slower than current computers that operate at speeds of billions of times per second, but the massive parallel processing of the intricately connected

brain enables it to perform feats of inference that are still far beyond the capabilities of computers.

A neural representation is not a static object like a word on paper or a street sign, but is rather a dynamic process involving ongoing change in many neurons and their interconnections. A population of neurons represents something by its pattern of firing. The brain is capable of a vast number of patterns: assuming that each neuron can fire 100 times per second, then the number of firing patterns of that duration is $(2^{100})^{100,000,000,000}$, a number far larger that the number of elementary particles in the universe, which is only about 10^{80}. I call this "Dickinson's theorem," after Emily Dickinson's beautiful poem "The Brain Is Wider Than the Sky." A pattern of activation in the brain constitutes a representation of something when there is a stable causal correlation between the firing of neurons in a population and the thing that is represented, such as an object or group of objects in the world (Eliasmith, 2005b; Parisien & Thagard, 2008). The claim that mental representations are patterns of firing in neural populations is a radical departure from everyday concepts and even from cognitive psychology until recently, but is increasingly supported by data acquired through experimental techniques such as brain scans and by rapidly developing theories about how brains work (e.g., Anderson, 2007; Smith & Kosslyn, 2007; Thagard 2010a).

Neural Mental Models

Demonstrating that neural representations can constitute mental models requires showing how they can have the same relational structure as what they represent, both statically and dynamically. Static mental models have spatial structure similar to what they represent, whereas dynamic mental models have similar temporal structure. Combined mental models capture both spatial and temporal structure, as when a person runs a mental movie that represents what happens in some complex visual situation such as two cars colliding.

The most straightforward kind of neural mental models are topographical sensory maps, for which Knudsen, du Lac, and Esterly (1987, p. 61) provide the following summary:

> The nervous system performs computations to process information that is biologically important. Some of these computations occur in maps—arrays of neurons in

which the tuning of neighboring neurons for a particular parameter value varies systematically. Computational maps transform the representation into a place-coded probability distribution that represents the computed values of parameters by sites of maximum relative activity. Numerous computational maps have been discovered, including visual maps of line orientation and direction of motion, auditory maps of amplitude spectrum and time interval, and motor maps of orienting movements.

The simplest example is the primary visual cortex, in which neighboring columns of neurons process information from neighboring small regions of visual space (Knudsen, du Lac & Esterly, 1987; Kaas, 1997). In this case, the spatial organization of the neurons corresponds systematically to the spatial organization of the world, in the same way that the location of major cities on a map of Brazil corresponds to the actual location of those cities.

Such topographic neural models are useful for basic perception, but they are not rich enough to support high-level kinds of reasoning such as the above "taller than" example. How populations of neurons can support such reasoning is still unknown, as brain scanning technologies do not have sufficient resolution to pin down neural activity in enough detail to inspire theoretical models of how high-level mental modeling can work. But let me try to extrapolate from current views on neural representation, particularly those of Eliasmith and Anderson (2003), to suggest how the brain might be able to make extra-topographic models of the world (see also Eliasmith, 2005b).

Neural populations can acquire the ability to encode features of the world as their firing activity becomes causally correlated with those features (*A* and *B* are causally correlated if they are statistically correlated as the result of causal interactions between *A* and *B*). Neural populations are also capable of encoding the activity of other neural populations, as the firing patterns of one population become causally correlated with the firing patterns of another population that feeds into it. If the input population is a topographic map, then the output population can become a more abstract representation of the features of the world, in two ways. The most basic retains some of the topographic structure of the input population, so that the output population is still a mental model of the world in that it shares some (but not all) relational structure with it. An even more abstract encoding is performed by an output neural population that captures key aspects of the encoding performed by the input population, but does so in a manner analogous to the way that language produces