BLACKWELL PHILOSOPHY GUIDES

The Blackwell Guide to the

Philosophy of Computing and Information

Edited by Luciano Floridi

Blackwell Publishing

The Blackwell Guide to

the Philosophy of Computing and Information

Blackwell Philosophy Guides

Series Editor: Steven M. Cahn, City University of New York Graduate School

Written by an international assembly of distinguished philosophers, the *Blackwell Philosophy Guides* create a groundbreaking student resource – a complete critical survey of the central themes and issues of philosophy today. Focusing and advancing key arguments throughout, each essay incorporates essential background material serving to clarify the history and logic of the relevant topic. Accordingly, these volumes will be a valuable resource for a broad range of students and readers, including professional philosophers.

1 The Blackwell Guide to Epistemology
Edited by John Greco and Ernest Sosa

2 The Blackwell Guide to Ethical Theory
Edited by Hugh LaFollette

3 The Blackwell Guide to the Modern Philosophers
Edited by Steven M. Emmanuel

4 The Blackwell Guide to Philosophical Logic
Edited by Lou Goble

5 The Blackwell Guide to Social and Political Philosophy
Edited by Robert L. Simon

6 The Blackwell Guide to Business Ethics
Edited by Norman E. Bowie

7 The Blackwell Guide to the Philosophy of Science
Edited by Peter Machamer and Michael Silberstein

8 The Blackwell Guide to Metaphysics
Edited by Richard M. Gale

9 The Blackwell Guide to the Philosophy of Education
Edited by Nigel Blake, Paul Smeyers, Richard Smith, and Paul Standish

10 The Blackwell Guide to Philosophy of Mind
Edited by Stephen P. Stich and Ted A. Warfield

11 The Blackwell Guide to the Philosophy of the Social Sciences
Edited by Stephen P. Turner and Paul A. Roth

12 The Blackwell Guide to Continental Philosophy
Edited by Robert C. Solomon and David Sherman

13 The Blackwell Guide to Ancient Philosophy
Edited by Christopher Shields

14 The Blackwell Guide to the Philosophy of Computing and Information
Edited by Luciano Floridi

15 The Blackwell Guide to Aesthetics
Edited by Peter Kivy

The Blackwell Guide to the Philosophy of Computing and Information

Edited by

Luciano Floridi

Blackwell Publishing

350 Main Street, Malden, MA 02148-5020, USA
108 Cowley Road, Oxford OX4 1JF, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2004 by Blackwell Publishing Ltd

Library of Congress Cataloging-in-Publication Data

The Blackwell guide to the philosophy of computing and information / edited by Luciano Floridi.
p. cm. — (Blackwell philosophy guides)
Includes bibliographical references and index.
ISBN 0-631-22918-3 (alk. paper) — ISBN 0-631-22919-1 (pbk. : alk. paper)
1. Computer science—Philosophy. 2. Information technology—Philosophy.
I. Floridi, Luciano, 1964– II. Series.

QA76.167.B53 2003
004′.01—dc21
2003045339

A catalogue record for this title is available from the British Library.

Set in 9/11.5pt Galliard
by Graphicraft Limited, Hong Kong
Printed and bound in the United Kingdom
by MPG Books Ltd, Bodmin, Cornwall

For further information on
Blackwell Publishing, visit our website:
http://www.blackwellpublishing.com

Contents

Notes on Contributors

Fred Adams is Professor of Cognitive Science and Philosophy, and Chair of the Department of Philosophy at the University of Delaware. He has also taught at Augustana College, Central Michigan University, Lawrence University, and the University of Wisconsin-Madison. A sample of his publications include the co-edited book *Reflections on Philosophy*, and authored or co-authored articles, "Cognitive Trying," "Causal Contents," "Fodorian Semantics," and "Vacuous Singular Terms."

G. Aldo Antonelli is Associate Professor of Logic and Philosophy of Science at the University of California, Irvine. He has worked in applications of logic to artificial intelligence and game theory, non-well-founded set theories, modal logic, and philosophy of language and mathematics.

Thierry Bardini is Associate Professor in the Communication Department at the University of Montréal. He holds a degree in agronomy (ENSA Montpellier, 1986) and a Ph.D. in sociology (Paris X, 1991). Since 1992 he has conducted research on the sociological history of computing, with a special emphasis on the genesis of personal computing.

Mark A. Bedau received his Ph.D. in philosophy from University of California at Berkeley in 1985. He is currently Professor of Philosophy and Humanities at Reed College in Portland, Oregon; Adjunct Professor of Systems Science at Portland State University; and Editor-in-chief of the journal *Artificial Life* (MIT Press). He has published extensively on both scientific and philosophical aspects of artificial life.

Cristina Bicchieri is Professor of Philosophy and Social and Decision Sciences at Carnegie Mellon University. She is the author of *Rationality and Coordination* (1993, 1997), and co-author of *The Dynamics of Norms* (1998), *The Logic of Strategy* (2000) and *Knowledge, Belief, and Strategic Interaction* (1992). She has published extensively in philosophy, AI, game theory, and the social sciences.

Jonathan Cohen is Assistant Professor of Philosophy at the University of California, San Diego. He is the author of several articles in philosophy of mind and philosophy of language. Much of his recent work has concerned the metaphysics of color.

Timothy Colburn has been Professor of Computer Science at the University of Minnesota-Duluth since 1988. Prior to that, he was principal research scientist for Honeywell Inc. He is the author of *Philosophy and Computer Science* (2000) and co-editor of *Program Verification: Fundamental Issues in Computer Science* (1993).

Wesley Cooper is Professor of Philosophy at the University of Alberta in Edmonton,

Alberta, Canada. He is the author of *The Unity of William James's Thought* (2002) and the chief administrator of Alberta MOO (http://www.arts.ualberta.ca:3000).

B. Jack Copeland is Professor of Philosophy at the University of Canterbury, New Zealand, and Director of the Turing Archive for the History of Computing. He works in mathematical and philosophical logic, cognitive science, and the history and foundations of computing, and has numerous articles in journals including *Journal of Philosophy*, *Mind*, *Analysis*, and *Scientific American*. He is author of *Artificial Intelligence: A Philosophical Introduction* (Blackwell, 1993, 2nd ed. forthcoming) and is currently writing and editing several books on Turing, including one on Turing's Automatic Computing Engine, and editing a volume entitled *Colossus: The First Electronic Computer*. His edited volume *Logic and Reality: Essays on the Legacy of Arthur Prior* appeared in 1996.

Roberto Cordeschi is Professor of Philosophy of Science at the University of Salerno, Italy. He is the author of several publications in the history of cybernetics and in the epistemological issues of cognitive science and artificial intelligence, including *The Discovery of the Artificial* (2002).

Charles Ess is Professor of Philosophy and Religion, and Director of the Center for Interdisciplinary Studies, Drury University. Ess has received awards for teaching excellence and scholarship, as well as a national award for his work in hypermedia. He has published in interdisciplinary ethics, hypertext and democratization, history of philosophy, feminist biblical studies, contemporary Continental philosophy, computer resources for humanists, and the interactions between culture, technology, and communication.

James H. Fetzer is McKnight Professor of Philosophy at the University of Minnesota, and teaches on its Duluth campus. The founding editor of the book series *Studies in Cognitive Systems* and of the journal *Minds and Machines*, he has published more than 20 books and 100 articles and reviews in the philosophy of science and on the theoretical foundations of computer science, artificial intelligence, and cognitive science.

Luciano Floridi is Associate Professor of Logic and Epistemology at the University of Bari, Italy, and Markle Foundation Fellow in Information Policy at the Oxford University in the UK. His publications include over 30 articles on the philosophy of computing and information and *Sextus Empiricus* (2002), *Philosophy and Computing – An Introduction* (1999), and *Scepticism and the Foundation of Epistemology* (1996). He was the consultant editor for the *Iter Italicum* on CD-ROM (1995) and for the *Routledge Encyclopedia of Philosophy on CD-ROM* (1998). He is the founding director of the Italian Web Site for Philosophy (www.swif.it).

Donald Gillies is Professor in the Philosophy Department of King's College, University of London. Since 1966 he has carried out research in the philosophy of science and mathematics. He has published 5 books, and edited a collection on "Revolutions in Mathematics." From the late 1980s, he has taken a particular interest in interactions between philosophy and AI, and has been involved in four interdisciplinary research projects in this area.

Patrick Grim is SUNY Distinguished Teaching Professor, State University of New York at Stony Brook. He is author of *The Incomplete Universe: Totality, Knowledge, and Truth* (1991), co-author of *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling* (with Gary Mar and Paul St. Denis, 1998), and founding co-author of 22 volumes of *The Philosopher's Annual*. He has published a variety of articles incorporating computer modeling in journals in philosophy, computer science, linguistics, and theoretical biology.

Deborah G. Johnson is the Anne Shirley Carter Olsson Professor of Applied Ethics in the Department of Technology, Culture, and Communication within the School of Engineering and Applied Science at the University of Virginia (Charlottesville, Virginia, USA). She specializes in ethical, social, and policy issues involving technology, especially issues involving computers and the internet.

Dominic McIver Lopes is Associate Professor of Philosophy at the University of British Columbia. He writes on issues at the intersection

of the philosophy of art and the philosophy of mind, and is the author of *Understanding Pictures* and co-editor of the *Routledge Companion to Aesthetics*. He is currently working on a book entitled *Live Wires: The Digital Arts*.

Klaus Mainzer is Professor of Philosophy of Science and Director of the Institute of Interdisciplinary Informatics (http://www.informatik.uni-augsburg.de/I3) at the University of Augsburg. He is president of the German Society of Complex Systems and Nonlinear Dynamics, author and editor of several books on philosophy of science, systems science, cognitive and computer science.

Brian P. McLaughlin is a Professor of Philosophy at Rutgers University and Chairperson of the Department. He is the author of numerous papers in the philosophy of mind, metaphysics, epistemology, and philosophical logic.

Carl Mitcham is Professor of Liberal Arts and International Studies at the Colorado School of Mines. His books include *Thinking through Technology: The Path between Engineering and Philosophy* (1994) and a co-edited (with Stephen H. Cutcliffe) volume on *Visions of STS: Counterpoints in Science, Technology, and Society Studies* (2001).

Barry Smith is Park Professor of Philosophy at the State University of New York at Buffalo and Director of the Institute for Formal Ontology and Medical Information Science at the University of Leipzig. He is editor of *The Monist* and author of numerous articles on ontology, the history of Austro-German philosophy, and related themes.

Derek Stanovsky is Director of Internet Studies at Appalachian State University and an Assistant Professor in the Department of Interdisciplinary Studies. His research interests include internet studies, feminist theory, and contemporary continental philosophy, and his articles have appeared in *The National Women's Studies Association Journal, Jouvert: A Journal of Postcolonial Studies*, and *Feminist Teacher*.

Eric Steinhart is Professor of Philosophy at William Paterson University. He works mainly on metaphysics and has written on possible worlds semantics for metaphors, the logical foundations of physical theory, theories of transfinite computation, and the nature of persons.

Paul Thagard is Professor of Philosophy and Director of the Cognitive Science Program at the University of Waterloo, Canada. His books include *Coherence in Thought and Action* (2000) and *How Scientists Explain Disease* (1999).

Alasdair Urquhart is Professor of Philosophy and Computer Science at the University of Toronto. He has published papers in non-classical logics, algebraic logic, and complexity theory, and is the editor of volume 4 of the *Collected Papers of Bertrand Russell*.

Graham White is Lecturer in Computer Science at Queen Mary, University of London. He has previously published on the history of logic and on the philosophy of common-sense reasoning; currently he is working on the logic of action and its computational implementation.

Preface

The information revolution has been changing the world profoundly, irreversibly and problematically for some time now, at a breathtaking pace and with an unprecedented scope. Every year, the world produces between 1 and 2 exabytes of data, that is, roughly 250 megabytes for every human being on earth (source: Lyman & Varian, online, see figure P1). An exabyte is approximately

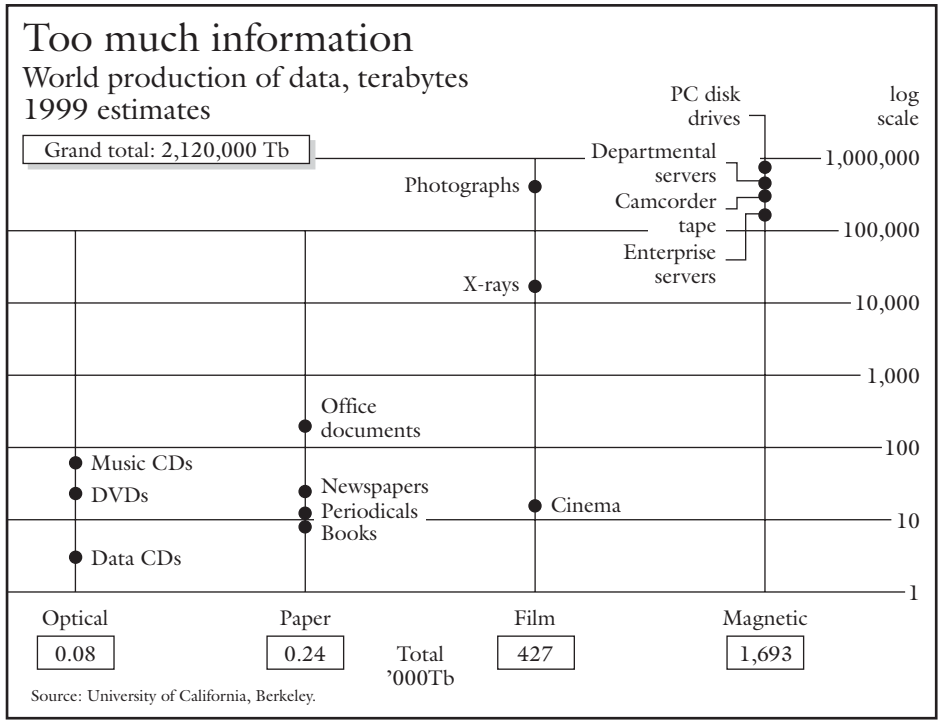

Figure P1

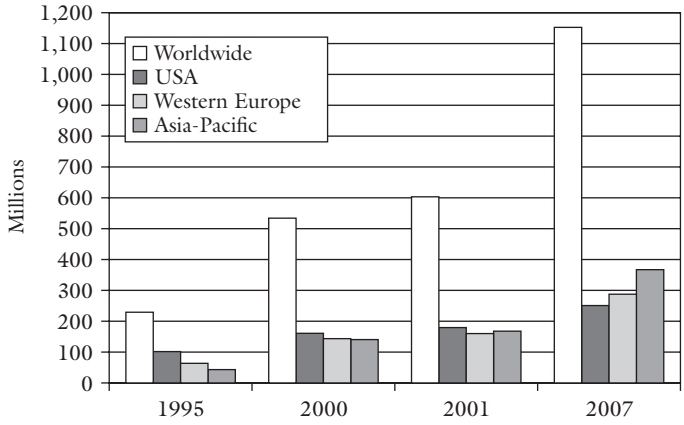

Figure P2: PCs in use worldwide
Source: Computer Industry Almanac Inc., <http://www.c-i-a.com/pr0302.htm>.

10^{18} bytes, or a billion times a billion bytes, the equivalent of about 20 billion copies of this *Guide*. It has taken the entire history of humanity to accumulate 12 exabytes of data. Stored on floppy disks, 12 exabytes of data would form a stack 24 million miles high. At the rate of growth measured in 1999, humanity will already have created the next 12 exabytes by the time this *Guide* is published.

To cope with exabytes of data, hundreds of millions of computing machines are employed every day. In 2001, the number of PCs in use worldwide reached 600M units (source: Computer Industry Almanac Inc., <http://www.c-i-a.com/pr0302.htm>, see figure P2). By the end of 2007, this number will have nearly doubled to over 1.15B PCs, at a compound annual growth of 11.4 percent. Of course, PCs are among the greatest sources of further exabytes.

Figures P1 and P2 give some quantitative substance to the trite remark that we live in a "databased society." They also show that the end of the information society, understood as the mature stabilization in the growth of quantity of data and number of computational machines, is not in sight.

The databased society has been brought about by the information revolution, whose main means have been first the PC and then the Web. "Datifying" the world and human society has created entirely new realities, made possible unprecedented phenomena and experiences, provided a wealth of extremely powerful tools and methodologies, raised a wide range of unique problems and conceptual issues, and opened up endless possibilities hitherto unimaginable.

Inevitably, the information revolution has also deeply affected what philosophers do, how they think about their problems, what problems they consider worth their attention, how they conceptualize their views, and even the vocabulary they use (see Bynum & Moor 1998 and 2002, Colburn 2000, Floridi 1999, and Mitcham & Huning 1986 for references). It has made possible new approaches and original investigations, posed or helped to identify unprecedented and crucial questions, and given new meaning to classic problems and traditional topics. In short, information-theoretic and computational research in philosophy has become increasingly innovative, fertile, and pervasive. It has already produced a wealth of interesting and important results. This *Guide* is the first systematic attempt to map this new and vitally important area of research. Owing to the novelty of the field, it is an exploration as much as an introduction.

As an introduction, the 26 chapters in this volume seek to provide a critical survey of the fundamental themes, problems, arguments, theories, and methodologies constituting the new field of *philosophy of computing and information* (PCI). The chapters are organized into seven sections. In Part I, four of the most crucial concepts in PCI, namely *computation*, *complexity*, *system*,

and *information* are analyzed. They are the four columns on which the other chapters are built, as it were. The following six parts are dedicated to specific areas: *the information society* (computer ethics; communication and interaction; cyberphilosophy and internet culture; and digital art); *mind and intelligence* (philosophy of AI and its critique; and computationalism, connectionism, and the philosophy of mind); *natural and artificial realities* (formal ontology; virtual reality; the physics of information; cybernetics; and artificial life); *language and knowledge* (meaning and information; knowledge and information; formal languages; and hypertext theory); *logic and probability* (nonmonotonic logic; probabilistic reasoning; and game theory); and, finally, *science, technology, and methodology* (computing in the philosophy of science; methodology of computer science; philosophy of IT; and computational modeling as a philosophical methodology). Each chapter has been planned as a freestanding introduction to its subject. For this purpose, the volume is further supported by an exhaustive glossary of technical terms, available online (http://www.blackwellpublishing.com/pci).

As an exploration, the *Guide* attempts to bring into a reasonable relation the many computational and informational issues with which philosophers have been engaged at least since the 1950s. The aim has been to identify a broad but clearly definable and well-delimited field where before there were many special problems and ideas whose interrelations were not always explicit or well understood. Each chapter is meant to provide not only a precise, clear, and accessible introduction but also a substantial and constructive contribution to the current debate.

Precisely because the *Guide* is also an exploration, the name given to the new field is somewhat tentative. Various labels have recently been suggested. Some follow fashionable terminology (e.g. "cyberphilosophy," "digital philosophy," "computational philosophy"), while the majority expresses specific theoretical orientations (e.g. "philosophy of computer science," "philosophy of computing/computation," "philosophy of AI," "philosophy and computers," "computing and philosophy," "philosophy of the artificial," "artificial epistemology," "android epistemology"). For this *Guide*, the philosophy editors at Blackwell and I agreed to use "philosophy of computing and information." PCI is a new but still very recognizable label, which we hope will serve both scholarly and marketing ends equally well. In the introductory chapter, entitled "What is the Philosophy of Information?," I offer an interpretation of the new informational paradigm in philosophy and argue that *philosophy of information* (PI) is conceptually a much more satisfactory name for it, because it identifies far more clearly what really lies at the heart of the new paradigm. But much as I hope that PI will become a useful label, I suspect that it would have been premature and somewhat obscure as the title for this volume. Since the chapter is meant to prepare the ground for the *Guide*, I thought it would be convenient to make it available on the Web free of charge (it can be found online at http://www.blackwellpublishing.com/pci). The reader may wish to consider that the project for the *Guide* was based on the hermeneutical frame outlined in that chapter.

Acknowledgments

Because of the innovative nature of the research area, working on this *Guide* has been very challenging. I relied on the patience and expertise of so many colleagues, friends, and family members that I wish to apologize in advance if I have forgotten to mention anyone below. Jim Moor was one of the first people with whom I discussed the project and I wish to thank him for his time, suggestions, and support. Jeff Dean, philosophy editor at Blackwell, has come close to instantiating the Platonic idea of editor, with many comments, ideas, suggestions, and the right kind of support. This *Guide* has been made possible also by his far-sighted faith in the project. Nick Bellorini, also editor at Blackwell, has been equally important in the second stage of the editorial project. I am also grateful to the two anonymous referees who provided constructive feedback. Many other colleagues, most of whom I have not met in real life, generously contributed to the shaping of the project by commenting on earlier drafts through several mailing lists, especially hopos-l@listserv.nd.edu, philinfo@yahoogroups.

com, philosl@liverpool.ac.uk, philosop@louisiana.edu, and silfs-l@list.cineca.it. I am grateful to the list moderators and to Bryan Alexander, Colin Allen, Leslie Burkholder, Rafael Capurro, Tony Chemero, Ron Chrisley, Stephen Clark, Anthony Dardis, M. G. Dastagir, Bob Di Falco, Soraj Hongladarom, Ronald Jump, Lou Marinoff, Ioan-Lucian Muntean, Eric Palmer, Mario Piazza, John Preston, Geoffrey Rockwell, Gino Roncaglia, Jeff Sanders, and Nelson Thompson. Unfortunately, for reasons of space, not all their suggestions could be followed in this context. Here are some of the topics left out or only marginally touched upon: information science as applied philosophy of information; social epistemology and the philosophy of information; visual thinking; pedagogical issues in PCI; the philosophy of information design and modeling; the philosophy of information economy; lambda calculus; linear logic; conditional reasoning; epistemic logic; deontic logic; temporal logic; the logic of action; fuzzy logic; situation logic; dynamic logic; common-sense reasoning and AI; causal reasoning, the hermeneutical interpretation of AI.

J. C. Beall, Jonathan Cohen, Gualtiero Piccinini, Luigi Dappiano, and Saul Fisher sent me useful feedback on an earlier draft of the Glossary.

Members of four research groups have played an influential role in the development of the project. I cannot thank all of them but I wish to acknowledge the help I have received from IACAP, the International Association for Computing and Philosophy, directed by Robert Cavalier (http://caae.phil.cmu.edu/caae/CAP/), with its meetings at Carnegie Mellon (CAP@CMU); INSEIT, the International Society for Ethics and Information Technology; the American Philosophical Association Committee on Philosophy and Computers (http://www.apa.udel.edu/apa/governance/committees/computers/); and the Epistemology and Computing Lab, directed by Mauro Di Giandomenico at the Philosophy Department of the University of Bari (http://www.ssscienza.uniba.it/index.html). I am also grateful to my College in Oxford, Wolfson, for the IT facilities that have made possible the organization of a website to support the editorial work (http://www.wolfson.ox.ac.uk/~floridi/blackwell/index.htm). During the editorial process, files were made available to all contributors through this website and I hope it will be possible to transform it into a permanent resource for the use of the *Guide*. The Programme in Comparative Media Law and Policy at Oxford University and its founding director Monroe Price greatly facilitated my work. Research for this project has been partly supported by a grant from the Coimbra Group, Pavia University, and I wish to thank Lorenzo Magnani, Director of the Computational Philosophy Lab (http://www.unipv.it/webphilos_lab/) and Mario Stefanelli, Director of the Medical Informatics Laboratory (http://dis.unipv.it/labs/labmed/home-e.html) for their support. Finally, I wish to thank all the contributors for bearing with me as chapters went through so many revisions; my father, for making me realize the obvious, namely the exploratory nature of this project; Paul Oldfield, for his copy-editing help; and my wife Kia, who not only implemented a wonderful life for our family, but also listened to me patiently when things were not working, provided many good solutions to problems in which I had entangled myself, and went as far as to read my contributions and comment carefully on their contents. The only thing she could not do was to take responsibility for any mistakes still remaining.

Luciano Floridi

References

Bynum, T. W. and Moor, J. H., eds. 1998. *The Digital Phoenix: How Computers are Changing Philosophy*. Malden, MA and Oxford: Blackwell.

—— and ——, eds. 2002. *CyberPhilosophy: The Intersection of Philosophy and Computing*. Malden, MA and Oxford: Blackwell.

Colburn, T. R. 2000. *Philosophy and Computer Science*. Armonk, NY and London: M. E. Sharpe.

Floridi, L. 1999. *Philosophy and Computing – An Introduction*. London and New York: Routledge.

Lyman, P. and Varian, H. R. (online). "How Much Information?," <http://www.sims.berkeley.edu/research/projects/how-much-info/index.html>.

Mitcham, C. and Huning, A., eds. 1986. *Philosophy and Technology II – Information Technology and Computers in Theory and Practice*. Dordrecht/Boston: Reidel.

Part I

Four Concepts

Chapter 1

Computation

B. Jack Copeland

The Birth of the Modern Computer

As everyone who can operate a personal computer knows, the way to make the machine perform some desired task is to open the appropriate program stored in the computer's memory. Life was not always so simple. The earliest large-scale electronic digital computers, the British Colossus (1943) and the American ENIAC (1945), did not store programs in memory (see Copeland 2001). To set up these computers for a fresh task, it was necessary to modify some of the machine's wiring, rerouting cables by hand and setting switches. The basic principle of the modern computer – the idea of controlling the machine's operations by means of a program of coded instructions stored in the computer's memory – was thought of by Alan Turing in 1935. His abstract "universal computing machine," soon known simply as the universal Turing machine (UTM), consists of a limitless memory, in which both data and instructions are stored, and a scanner that moves back and forth through the memory, symbol by symbol, reading what it finds and writing further symbols. By inserting different programs into the memory, the machine is made to carry out different computations.

Turing's idea of a universal stored-program computing machine was promulgated in the US by John von Neumann and in the UK by Max Newman, the two mathematicians who were by and large responsible for placing Turing's abstract universal machine into the hands of electronic engineers (Copeland 2001). By 1945, several groups in both countries had embarked on creating a universal Turing machine in hardware. The race to get the first electronic stored-program computer up and running was won by Manchester University where, in Newman's Computing Machine Laboratory, the "Manchester Baby" ran its first program on June 21, 1948. By 1951, electronic stored-program computers had begun to arrive in the marketplace. The first model to go on sale was the Ferranti Mark I, the production version of the Manchester computer (built by the Manchester firm Ferranti Ltd.). Nine of the Ferranti machines were sold, in Britain, Canada, Holland, and Italy, the first being installed at Manchester University in February 1951. In the US, the Computer Corporation sold its first UNIVAC later the same year. The LEO computer also made its debut in 1951; LEO was a commercial version of the prototype EDSAC machine, which at Cambridge University in 1949 had become the second stored-program electronic computer to function. In 1953 came the IBM 701, the company's first mass-produced stored-program electronic computer (strongly influenced by von Neumann's prototype IAS computer, which was working at

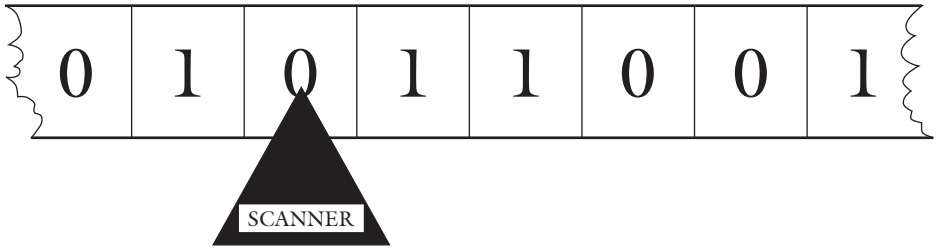

Figure 1.1: A Turing machine

Princeton University by the summer of 1951). A new era had begun.

Turing introduced his abstract Turing machines in a famous article entitled "On Computable Numbers, with an Application to the Entscheidungsproblem" (published in 1936). Turing referred to his abstract machines simply as "computing machines" – the American logician Alonzo Church dubbed them "Turing machines" (Church 1937: 43). "On Computable Numbers" pioneered the theory of computation and is regarded as the founding publication of the modern science of computing. In addition, Turing charted areas of mathematics lying beyond the reach of the UTM. He showed that not all precisely-stated mathematical problems can be solved by a Turing machine. One of them is the *Entscheidungsproblem* – "decision problem" – described below. This discovery wreaked havoc with received mathematical and philosophical opinion. Turing's work – together with contemporaneous work by Church (1936a, 1936b) – initiated the important branch of mathematical logic that investigates and codifies problems "too hard" to be solvable by Turing machine. In a single article, Turing ushered in both the modern computer and the mathematical study of the uncomputable.

What is a Turing Machine?

A Turing machine consists of a limitless memory and a scanner that moves back and forth through the memory, symbol by symbol, reading what it finds and writing further symbols. The memory consists of a tape divided into squares. Each square may be blank or may bear a single symbol, "0" or "1," for example, or some other symbol taken from a finite alphabet. The scanner is able to examine only one square of tape at a time (the "scanned square"). (See figure 1.1.) The tape is the machine's general-purpose storage medium, serving as the vehicle for input and output, and as a working memory for storing the results of intermediate steps of the computation. The tape may also contain a program of instructions. The input that is inscribed on the tape before the computation starts must consist of a finite number of symbols. However, the tape itself is of unbounded length – since Turing's aim was to show that there are tasks which these machines are unable to perform, even given unlimited working memory and unlimited time. (A Turing machine with a tape of fixed finite length is called a *finite state automaton*. The theory of finite state automata is not covered in this chapter. An introduction may be found in Sipser 1997.)

The Basic Operations of a Turing Machine

Each Turing machine has the same small repertoire of *basic* (or "atomic") operations. These are logically simple. The scanner contains mechanisms that enable it to *erase* the symbol on the scanned square, to *write* a symbol on the scanned square (first erasing any existing symbol), and to *shift position* one square to the left or right. Complexity of operation is achieved by chaining together large numbers of these simple

Table 1.1

State	*Scanned square*	*Operations*	*Next state*
a	blank	P[0], R	**b**
b	blank	R	**c**
c	blank	P[1], R	**d**
d	blank	R	**a**

basic actions. The scanner will *halt* if instructed to do so, i.e. will cease work, coming to rest on some particular square, for example the square containing the output (or if the output consists of a string of several digits, then on the square containing the left-most digit of the output, say).

In addition to the operations just mentioned, *erase*, *write*, *shift*, and *halt*, the scanner is able to *change state*. A device within the scanner is capable of adopting a number of different positions. This device may be conceptualized as consisting of a dial with a finite number of positions, labeled "a," "b," "c," etc. Each of these positions counts as a different state, and changing state amounts to shifting the dial's pointer from one labeled position to another. The device functions as a simple memory. As Turing said, by altering its state the "machine can effectively remember some of the symbols which it has 'seen' (scanned) previously" (1936: 231). For example, a dial with two positions can be used to keep a record of which binary digit, 0 or 1, is present on the square that the scanner has just vacated. If a square might also be blank, then a dial with three positions is required.

Commercially available computers are hard-wired to perform basic operations considerably more sophisticated than those of a Turing machine – add, multiply, decrement, store-at-address, branch, and so forth. The precise list of basic operations varies from manufacturer to manufacturer. It is a remarkable fact that none of these computers can out-compute the UTM. Despite the austere simplicity of Turing's machines, they are capable of computing anything that any computer on the market can compute. Indeed, because they are abstract machines, they are capable of computations that no "real" computer could perform.

Example of a Turing machine

The following simple example is from "On Computable Numbers" (Turing 1936: 233). The machine – call it **M** – starts work with a blank tape. The tape is endless. The problem is to set up the machine so that if the scanner is positioned over any square of the tape and the machine set in motion, it will print alternating binary digits on the tape, 0 1 0 1 0 1 . . . , working to the right from its starting place, leaving a blank square in between each digit. In order to do its work **M** makes use of four states labeled "**a**," "**b**," "**c**," and "**d**." **M** is in state **a** when it starts work. The operations that **M** is to perform can be set out by means of a table with four columns (see table 1.1). "R" abbreviates the instruction "shift right one square," "P[0]" abbreviates "print 0 on the scanned square," and likewise "P[1]." The top line of table 1.1 reads: if you are in state **a** and the square you are scanning is blank, then print 0 on the scanned square, shift right one square, and go into state **b**. A machine acting in accordance with this table of instructions – or program – toils endlessly on, printing the desired sequence of digits while leaving alternate squares blank.

Turing did not explain how it is to be brought about that the machine acts in accordance with the instructions. There was no need. Turing's machines are abstractions and it is not necessary to propose any specific mechanism for causing the machine to follow the instructions. However, for purposes of visualization, one might imagine the scanner to be accompanied by a bank of switches and plugs resembling an old-fashioned telephone switchboard. Arranging the plugs and setting the switches in a certain way causes the machine to act in accordance

with the instructions in table 1.1. Other ways of setting up the "switchboard" cause the machine to act in accordance with other tables of instructions.

The universal Turing machine

The UTM has a single, fixed table of instructions, which we may imagine to have been set into the machine by way of the switchboard-like arrangement just mentioned. Operating in accordance with this table of instructions, the UTM is able to carry out *any* task for which a Turing-machine instruction table can be written. The trick is to place an instruction table for carrying out the desired task onto the tape of the universal machine, the first line of the table occupying the first so many squares of the tape, the second line the next so many squares, and so on. The UTM reads the instructions and carries them out on its tape. This ingenious idea is fundamental to computer science. The universal Turing machine is in concept the stored-program digital computer.

Turing's greatest contributions to the development of the modern computer were:

- The idea of controlling the function of the computing machine by storing a program of (symbolically or numerically encoded) instructions in the machine's memory.
- His proof that, by this means, a *single* machine of *fixed structure* is able to carry out every computation that can be carried out by any Turing machine whatsoever.

Human Computation

When Turing wrote "On Computable Numbers," a computer was not a machine at all, but a human being – a mathematical assistant who calculated by rote, in accordance with some "effective method" supplied by an overseer prior to the calculation. A paper-and-pencil method is said to be effective, in the mathematical sense, if it (a) demands no insight or ingenuity from the human carrying it out, and (b) produces the correct answer in a finite number of steps. (An example of an effective method well-known among philosophers is the truth table test for tautologousness.) Many thousands of human computers were employed in business, government, and research establishments, doing some of the sorts of calculating work that nowadays is performed by electronic computers. Like filing clerks, computers might have little detailed knowledge of the end to which their work was directed.

The term "computing machine" was used to refer to calculating machines that mechanized elements of the human computer's work. These were in effect homunculi, calculating more quickly than an unassisted human computer, but doing nothing that could not in principle be done by a human clerk working effectively. Early computing machines were somewhat like today's nonprogrammable hand-calculators: they were not automatic, and each step – each addition, division, and so on – was initiated manually by the human operator. For a complex calculation, several dozen human computers might be required, each equipped with a desk-top computing machine. By the 1940s, however, the scale of some calculations required by physicists and engineers had become so great that the work could not easily be done in a reasonable time by even a roomful of human computers with desk-top computing machines. The need to develop high-speed, large-scale, automatic computing machinery was pressing.

In the late 1940s and early 1950s, with the advent of electronic computing machines, the phrase "computing machine" gave way gradually to "computer." During the brief period in which the old and new meanings of "computer" co-existed, the prefix "electronic" or "digital" would usually be used in order to distinguish machine from human. As Turing stated, the new electronic machines were "intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner" (Turing 1950: 1). Main-frames, laptops, pocket calculators, palm-pilots – all carry out work that a human rote-worker could do, if he or she

worked long enough, and had a plentiful enough supply of paper and pencils.

The Turing machine is an idealization of the human computer (Turing 1936: 231). Wittgenstein put this point in a striking way:

> Turing's "Machines." These machines are *humans* who calculate. (Wittgenstein 1980: §1096)

It was not, of course, some deficiency of imagination that led Turing to model his logical computing machines on what can be achieved by a human being working effectively. The purpose for which he introduced them demanded it. The Turing machine played a key role in his demonstration that there are mathematical tasks which *cannot* be carried out by means of an effective method.

The Church–Turing Thesis

The concept of an effective method is an informal one. Attempts such as the above to explain what counts as an effective method are not rigorous, since the requirement that the method demand neither insight nor ingenuity is left unexplicated. One of Turing's leading achievements – and this was a large first step in the development of the mathematical theory of computation – was to propose a rigorously defined expression with which the informal expression "by means of an effective method" might be replaced. The rigorously defined expression, of course, is "by means of a Turing machine." The importance of Turing's proposal is this: if the proposal is correct, then talk about the existence and non-existence of effective methods can be replaced throughout mathematics and logic by talk about the existence or non-existence of Turing machine programs. For instance, one can establish that there is no effective method at all for doing such-and-such a thing by proving that no Turing machine can do the thing in question.

Turing's proposal is encapsulated in the *Church–Turing thesis*, also known simply as *Turing's thesis*:

> The UTM is able to perform any calculation that any human computer can carry out.

An equivalent way of stating the thesis is:

> Any effective – or *mechanical* – method can be carried out by the UTM.

("Mechanical" is a term of art in mathematics and logic. It does not carry its everyday meaning, being in its technical sense simply a synonym for "effective.") Notice that the converse of the thesis – any problem-solving method that can be carried out by the UTM is effective – is obviously true, since a human being can, in principle, work through any Turing-machine program, obeying the instructions ("in principle" because we have to assume that the human does not go crazy with boredom, or die of old age, or use up every sheet of paper in the universe).

Church independently proposed a different way of replacing talk about effective methods with formally precise language (Church 1936a). Turing remarked that his own way of proceeding was "possibly more convincing" (1937: 153); Church acknowledged the point, saying that Turing's concept of computation by Turing machine "has the advantage of making the identification with effectiveness . . . evident immediately" (Church 1937: 43).

The name "Church–Turing thesis," now standard, seems to have been introduced by Kleene, with a flourish of bias in favor of his mentor Church (Kleene 1967: 232):

> Turing's and Church's theses are equivalent. We shall usually refer to them both as *Church's thesis*, or in connection with that one of its . . . versions which deals with "Turing machines" as *the Church–Turing thesis.*

Soon ample evidence amassed for the Church–Turing thesis. (A survey is given in chs. 12 and 13 of Kleene 1952.) Before long it was (as Turing put it) "agreed amongst logicians" that his proposal gives the "correct accurate rendering" of talk about effective methods (Turing 1948: 7). (Nevertheless, there have been occasional dissenting voices over the years; for example Kalmár 1959 and Péter 1959.)

Beyond the Universal Turing Machine

Computable and uncomputable numbers

Turing calls any number that can be written out by a Turing machine a *computable* number. That is, a number is computable, in Turing's sense, if and only if there is a Turing machine that calculates each digit of the number's decimal representation, in sequence. π, for example, is a computable number. A suitably programmed Turing machine will spend all eternity writing out the decimal representation of π digit by digit, 3.14159 . . .

Straight off, one might expect it to be the case that *every* number that *has* a decimal representation (that is to say, every real number) is computable. For what could prevent there being, for any particular number, a Turing machine that "churns out" that number's decimal representation digit by digit? However, Turing proved that not every real number is computable. In fact, computable numbers are relatively scarce among the real numbers. There are only *countably* many computable numbers, because there are only countably many different Turing-machine programs (instruction tables). (A collection of things is countable if and only if either the collection is finite or its members can be put into a one-to-one correspondence with the integers, 1, 2, 3, ...) As Georg Cantor proved in 1874, there are *uncountably* many real numbers – in other words, there are more real numbers than integers. There are literally not enough Turing-machine programs to go around in order for every real number to be computable.

The printing problem and the halting problem

Turing described a number of mathematical problems that cannot be solved by Turing machine. One is the *printing problem*. Some programs print "0" at some stage in their computations; all the remaining programs never print "0." The printing problem is the problem of deciding, given any arbitrarily selected program, into which of these two categories it falls. Turing showed that this problem cannot be solved by the UTM.

The *halting problem* (Davis 1958) is another example of a problem that cannot be solved by the UTM (although not one explicitly considered by Turing). This is the problem of determining, given any arbitrary Turing machine, whether or not the machine will eventually halt when started on a blank tape. The machine shown in table 1.1 is rather obviously one of those that never halts – but in other cases it is definitely not obvious from a machine's table whether or not it halts. And, of course, simply watching the machine run (or a simulation of the machine) is of no help at all, for what can be concluded if after a week or a year the machine has not halted? If the machine does eventually halt, a watching human – or Turing machine – will sooner or later find this out; but in the case of a machine that has not yet halted, there is no effective method for deciding whether or not it is going to halt.

The halting function

A *function* is a mapping from "arguments" (or inputs) to "values" (or outputs). For example, addition (+) is a function that maps pairs of numbers to single numbers: the value of the function + for the pair of arguments 5, 7 is the number 12. The squaring function maps single numbers to single numbers: e.g. the value of n^2 for the argument 3 is 9.

A function is said to be *computable by Turing machine* if some Turing machine will take in arguments of the function (or pairs of arguments, etc.) and, after carrying out some finite number of basic operations, produce the corresponding value – and, moreover, will do this *no matter which* argument of the function is presented. For example, addition over the integers is computable by Turing machine, since a Turing machine can be set up so that whenever two integers are inscribed on its tape (in binary notation, say), the machine will output their sum.

The *halting function* is as follows. Assume the Turing machines to be ordered in some way, so that we may speak of the first machine in the

ordering, the second, and so on. (There are various standard ways of accomplishing this ordering, e.g. in terms of the number of symbols in each machine's instruction table.) The arguments of the halting function are simply 1, 2, 3, ... (Like the squaring function, the halting function takes single arguments.) The value of the halting function for any argument n is 1 if the n^{th} Turing machine in the ordering eventually halts when started on a blank tape, and is 0 if the n^{th} machine runs on forever (as would, for example, a Turing machine programmed to produce in succession the digits of the decimal representation of π).

The theorem that the UTM cannot solve the halting problem is often expressed in terms of the halting function.

Halting theorem: The halting function is not computable by Turing machine.

The Entscheidungsproblem

The *Entscheidungsproblem*, or decision problem, was Turing's principal quarry in "On Computable Numbers." The decision problem was brought to the fore of mathematics by the German mathematician David Hilbert (who in a lecture given in Paris in 1900 set the agenda for much of twentieth-century mathematics). Hilbert and his followers held that mathematicians should seek to express mathematics in the form of a complete, consistent, decidable formal system – a system expressing "the entire thought-content of mathematics in a uniform way" (Hilbert 1927: 475). The project of formulating mathematics in this way became known as the "Hilbert program."

A consistent system is one that contains no contradictions; a complete system one in which every true mathematical statement is provable. "Decidable" means that there is an effective method for telling, of each mathematical statement, whether or not the statement is provable in the system. A complete, consistent, decidable system would banish ignorance from mathematics. Given any mathematical statement, one would be able to tell whether the statement is true or false by deciding whether or not it is provable in the system. Hilbert famously declared in his Paris lecture: "in mathematics there is no *ignorabimus*" (there is no *we shall not know*) (Hilbert 1902: 445).

It is important that the system expressing the "whole thought content of mathematics" be consistent. An inconsistent system – a system containing contradictions – is worthless, since *any* statement whatsoever, true or false, can be derived from a contradiction by simple logical steps. So in an inconsistent system, absurdities such as $0 = 1$ and $6 \neq 6$ are provable. An inconsistent system would indeed contain all true mathematical statements – would be complete, in other words – but would in addition also contain all false mathematical statements.

If ignorance is to be banished absolutely, the system must be decidable. An undecidable system might on occasion leave us in ignorance. Only if the mathematical system were decidable could we be confident of always being able to tell whether or not any given statement is provable. Unfortunately for the Hilbert program, however, it became clear that most interesting mathematical systems are, if consistent, incomplete and undecidable.

In 1931 Gödel showed that Hilbert's ideal is impossible to satisfy, even in the case of simple arithmetic. He proved that the system called Peano arithmetic is, if consistent, incomplete. This is known as Gödel's *first incompleteness theorem*. (Gödel later generalized this result, pointing out that "due to A. M. Turing's work, a precise and unquestionably adequate definition of the general concept of formal system can now be given," with the consequence that incompleteness can "be proved rigorously for *every* consistent formal system containing a certain amount of finitary number theory" (Gödel 1965: 71).) Gödel had shown that no matter how hard mathematicians might try to construct the all-encompassing formal system envisaged by Hilbert, the product of their labors would, if consistent, inevitably be incomplete. As Hermann Weyl – one of Hilbert's greatest pupils – observed, this was nothing less than "a catastrophe" for the Hilbert program (Weyl 1944: 644).

Gödel's theorem does not mention decidability. This aspect was addressed by Turing and by Church. Each showed, working independently, that no consistent formal system of arithmetic is

decidable. They showed this by proving that not even the weaker, purely logical system presupposed by any formal system of arithmetic and called the *first-order predicate calculus* is decidable. Turing's way of proving that the first-order predicate calculus is undecidable involved the printing problem. He showed that if a Turing machine could tell, of any given statement, whether or not the statement is provable in the first-order predicate calculus, then a Turing machine could tell, of any given Turing machine, whether or not it ever prints "0." Since, as he had already established, no Turing machine can do the latter, it follows that no Turing machine can do the former. The final step of the argument is to apply Turing's thesis: if no Turing machine can perform the task in question, then there is no effective method for performing it. The Hilbertian dream lay in total ruin.

Poor news though Turing's and Church's result was for the Hilbert school, it was welcome news in other quarters, for a reason that Hilbert's illustrious pupil von Neumann had given in 1927 (von Neumann 1927: 12):

> If undecidability were to fail then mathematics, in today's sense, would cease to exist; its place would be taken by a completely mechanical rule, with the aid of which any man would be able to decide, of any given statement, whether the statement can be proven or not.

In a similar vein, the Cambridge mathematician G. H. Hardy said in a lecture in 1928 (Hardy 1929: 16):

> if there were . . . a mechanical set of rules for the solution of all mathematical problems . . . our activities as mathematicians would come to an end.

The next section is based on Copeland 1996.

Misunderstandings of the Church–Turing Thesis: The Limits of Machines

A myth has arisen concerning Turing's work, namely that he gave a treatment of the limits of mechanism, and established a fundamental result to the effect that the UTM can simulate the behavior of *any* machine. The myth has passed into the philosophy of mind, theoretical psychology, cognitive science, Artificial Intelligence, and Artificial Life, generally to pernicious effect. For example, the *Oxford Companion to the Mind* states: "Turing showed that his very simple machine . . . can specify the steps required for the solution of any problem that can be solved by instructions, explicitly stated rules, or procedures" (Gregory 1987: 784). Dennett maintains that "Turing had proven – and this is probably his greatest contribution – that his Universal Turing machine can compute any function that any computer, with any architecture, can compute" (1991: 215); also that every "task for which there is a clear recipe composed of simple steps can be performed by a very simple computer, a universal Turing machine, the universal recipe-follower" (1978: xviii). Paul and Patricia Churchland assert that Turing's "results entail something remarkable, namely that a standard digital computer, given only the right program, a large enough memory and sufficient time, can compute *any* rule-governed input–output function. That is, it can display any systematic pattern of responses to the environment whatsoever" (1990: 26). Even Turing's biographer, Hodges, has endorsed the myth:

> Alan had . . . discovered something almost . . . miraculous, the idea of a universal machine that could take over the work of *any* machine. (Hodges 1992: 109)

Turing did not show that his machines can solve any problem that can be solved "by instructions, explicitly stated rules, or procedures," and nor did he prove that the UTM "can compute any function that any computer, with any architecture, can compute" or perform any "task for which there is a clear recipe composed of simple steps." As previously explained, what he proved is that the UTM can carry out any task that any *Turing machine* can carry out. Each of the claims just quoted says considerably more than this.

If what the Churchlands assert were true, then the view that psychology must be capable of being expressed in standard computational terms

would be secure (as would a number of other controversial claims). But Turing had no result entailing that "a standard digital computer . . . can compute *any* rule-governed input–output function." What he did have was a result entailing the exact opposite. The theorem that no Turing machine can decide the predicate calculus entails that there are rule-governed input–output functions that no Turing machine is able to compute – for example, the function whose output is 1 whenever the input is a statement that is provable in the predicate calculus, and is 0 for all other inputs. There are certainly possible patterns of responses to the environment, perfectly systematic patterns, that no Turing machine can display. One is the pattern of responses just described. The halting function is a mathematical characterization of another such pattern.

Distant cousins of the Church–Turing thesis

As has already been emphasized, the Church–Turing thesis concerns the extent of effective methods. Putting this another way (and ignoring contingencies such as boredom, death, or insufficiency of paper), the thesis concerns what a *human being* can achieve when working by rote with paper and pencil. The thesis carries no implication concerning the extent of what *machines* are capable of achieving (even digital machines acting in accordance with "explicitly stated rules"). For among a machine's repertoire of basic operations, there may be those that no human working by rote with paper and pencil can perform.

Essentially, then, the Church–Turing thesis says that no human computer, or machine that mimics a human computer, can out-compute the UTM. However, a variety of other propositions, very different from this, are from time to time called the Church–Turing thesis (or Church's thesis), sometimes but not always with accompanying hedges such as "strong form" and "physical version." Some examples from the recent literature are given below. This loosening of established terminology is unfortunate, and can easily lead to misunderstandings. In what follows I use the expression "Church–Turing thesis properly so called" for the proposition that Turing and Church themselves endorsed.

> [C]onnectionist models . . . may possibly even challenge the strong construal of Church's Thesis as the claim that the class of well-defined computations is exhausted by those of Turing machines. (Smolensky 1988: 3)

> Church–Turing thesis: If there is a well defined procedure for manipulating symbols, then a Turing machine can be designed to do the procedure. (Henry 1993: 149)

> [I]t is difficult to see how any language that could actually be run on a physical computer could do more than Fortran can do. The idea that there is no such language is called Church's thesis. (Geroch & Hartle 1986: 539)

> The first aspect that we examine of Church's Thesis . . . [w]e can formulate, more precisely: The behaviour of any discrete physical system evolving according to local mechanical laws is recursive. (Odifreddi 1989: 107)

> I can now state the physical version of the Church–Turing principle: "Every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means." This formulation is both better defined and more physical than Turing's own way of expressing it. (Deutsch 1985: 99)

> That there exists a most general formulation of machine and that it leads to a unique set of input–output functions has come to be called *Church's thesis.* (Newell 1980: 150)

The maximality thesis

It is important to distinguish between the Church–Turing thesis properly so called and what I call the "maximality thesis" (Copeland 2000). (Among the few writers to distinguish explicitly between Turing's thesis and stronger propositions along the lines of the maximality thesis are Gandy 1980 and Sieg 1994.)

A machine m is said to be able to *generate* a certain function if m can be set up so that if m is

presented with any of the function's arguments, *m* will carry out some finite number of atomic processing steps at the end of which *m* produces the corresponding value of the function (*mutatis mutandis* in the case of functions that, like addition, demand more than one argument).

> **Maximality Thesis**: All functions that can be generated by machines (working on finite input in accordance with a finite program of instructions) are computable by Turing machine.

The maximality thesis ("thesis M") admits of two interpretations, according to whether the phrase "can be generated by machine" is taken in the this-worldly sense of "can be generated by a machine that conforms to the physical laws (if not to the resource constraints) of the actual world," or in a sense that abstracts from whether or not the envisaged machine could exist in the actual world. Under the latter interpretation, thesis M is false. It is straightforward to describe abstract machines that generate functions that cannot be generated by the UTM (see e.g. Abramson 1971, Copeland 2000, Copeland & Proudfoot 2000, Stewart 1991). Such machines are termed "hypercomputers" in Copeland and Proudfoot (1999a).

It is an open empirical question whether or not the this-worldly version of thesis M is true. Speculation that there may be physical processes – and so, potentially, machine-operations – whose behavior conforms to functions not computable by Turing machine stretches back over at least five decades. (Copeland & Sylvan 1999 is a survey; see also Copeland & Proudfoot 1999b.)

A source of potential misunderstanding about the limits of machines lies in the difference between the technical and everyday meanings of the word "mechanical." As previously remarked, in technical contexts "mechanical" and "effective" are often used interchangeably. (Gandy 1988 outlines the history of this usage of the word "mechanical.") For example:

> Turing proposed that a certain class of abstract machines could perform any "mechanical" computing procedure. (Mendelson 1964: 229)

Understood correctly, this remark attributes to Turing not a thesis concerning the limits of what can be achieved by machine but the Church–Turing thesis properly so called.

The technical usage of "mechanical" tends to obscure the possibility that there may be machines, or biological organs, that generate (or compute, in a broad sense) functions that cannot be computed by Turing machine. For the question "Can a machine execute a procedure that is not mechanical?" may appear self-answering, yet this is precisely what is asked if thesis M is questioned.

In the technical literature, the word "computable" is often tied by definition to effectiveness: a function is said to be computable if and only if there is an effective method for determining its values. The Church–Turing thesis then becomes:

> Every computable function can be computed by Turing machine.

Corollaries such as the following are sometimes stated:

> [C]ertain functions are uncomputable in an absolute sense: uncomputable even by [Turing machine], and, therefore, uncomputable by any past, present, or future real machine. (Boolos & Jeffrey 1980: 55)

When understood in the sense in which it is intended, this remark is perfectly true. However, to a casual reader of the technical literature, such statements may appear to say more than they in fact do.

Of course, the decision to tie the term "computable" and its cognates to the concept of effectiveness does not settle the truth-value of thesis M. Those who abide by this terminological decision will not describe a machine that falsifies thesis M as *computing* the function that it generates.

Putnam is one of the few writers on the philosophy of mind to question the proposition that Turing machines provide a maximally general formulation of the notion of machine:

> [M]aterialists are committed to the view that a human being is – at least metaphorically – a machine. It is understandable that the notion of a Turing machine might be seen as just a

> way of making this materialist idea precise. Understandable, but hardly well thought out. The problem is the following: a "machine" in the sense of a physical system obeying the laws of Newtonian physics need not be a Turing machine. (Putnam 1992: 4)

The Church–Turing fallacy

To commit what I call the *Church–Turing fallacy* (Copeland 2000, 1998) is to believe that the Church–Turing thesis, or some formal or semi-formal result established by Turing or Church, secures the following proposition:

> If the mind–brain is a machine, then the Turing-machine computable functions provide sufficient mathematical resources for a full account of human cognition.

Perhaps some who commit this fallacy are misled purely by the terminological practice already mentioned, whereby a thesis concerning which there is little real doubt, the Church–Turing thesis properly so called, and a nexus of different theses, some of unknown truth-value, are all referred to as Church's thesis or the Church–Turing thesis.

The Church–Turing fallacy has led to some remarkable claims in the foundations of psychology. For example, one frequently encounters the view that psychology *must* be capable of being expressed ultimately in terms of the Turing machine (e.g. Fodor 1981: 130; Boden 1988: 259). To anyone in the grip of the Church–Turing fallacy, conceptual space will seem to contain no room for mechanical models of the mind–brain that are not equivalent to a Turing machine. Yet it is certainly possible that psychology will find the need to employ models of human cognition that transcend Turing machines (see Chapter 10, COMPUTATIONALISM, CONNECTIONISM, AND THE PHILOSOPHY OF MIND).

The simulation fallacy

A closely related error, unfortunately also common in modern writing on computation and the brain, is to hold that Turing's results somehow entail that the brain, and indeed any biological or physical system whatever, can be *simulated* by a Turing machine. For example, the entry on Turing in *A Companion to the Philosophy of Mind* contains the following claims: "we can depend on there being a Turing machine that captures the functional relations of the brain," for so long as "these relations between input and output are functionally well-behaved enough to be describable by . . . mathematical relationships . . . we know that some specific version of a Turing machine will be able to mimic them" (Guttenplan 1994: 595). Even Dreyfus, in the course of *criticizing* the view that "man is a Turing machine," succumbs to the belief that it is a "fundamental truth that every form of 'information processing' (even those which *in practice* can only be carried out on an 'analogue computer') must *in principle* be simulable on a [Turing machine]" (1992: 195).

Searle writes in a similar fashion:

> If the question ["Is consciousness computable?"] asks "Is there some level of description at which conscious processes and their correlated brain processes can be simulated [by a Turing machine]?" the answer is trivially yes. Anything that can be described as a precise series of steps can be simulated [by a Turing machine]. (Searle 1997: 87)

> Can the operations of the brain be simulated on a digital computer? . . . The answer seems to me . . . demonstrably "Yes" . . . That is, naturally interpreted, the question means: Is there some description of the brain such that under that description you could do a computational simulation of the operations of the brain. But given Church's thesis that anything that can be given a precise enough characterization as a set of steps can be simulated on a digital computer, it follows trivially that the question has an affirmative answer. (Searle 1992: 200)

Church's thesis properly so called does *not* say that anything that can be described as a precise series of of steps can be simulated by Turing machine.

Similarly, Johnson-Laird and the Churchlands argue:

> If you assume that [consciousness] is scientifically explicable . . . [and] [g]ranted that the [Church–Turing] thesis is correct, then the final dichotomy rests on Craik's functionalism. If you believe [functionalism] to be false . . . then presumably you hold that consciousness could be modelled in a computer program in the same way that, say, the weather can be modelled . . . If you accept functionalism, however, then you should believe that consciousness is a computational process. (Johnson-Laird 1987: 252)

> Church's Thesis says that whatever is computable is Turing computable. Assuming, with some safety, that what the mind-brain does is computable, then it can in principle be simulated by a computer. (Churchland & Churchland 1983: 6)

As previously mentioned, the Churchlands believe, incorrectly, that Turing's "results entail . . . that a standard digital computer, given only the right program, a large enough memory and sufficient time, can . . . display any systematic pattern of responses to the environment whatsoever" (1990: 26). This no doubt explains why they think they can assume "with some safety" that what the mind–brain does is computable, for on their understanding of matters, this is to assume only that the mind–brain is characterized by a "rule-governed" (1990: 26) input–output function.

The Church–Turing thesis properly so called does not entail that the brain (or the mind, or consciousness) can be simulated by a Turing machine, not even in conjunction with the belief that the brain (or mind, etc.) is scientifically explicable, or exhibits a systematic pattern of responses to the environment, or is "rule-governed" (etc.). Each of the authors quoted seems to be assuming the truth of a close relative of thesis M, which I call "thesis S" (Copeland 2000).

> **Thesis S**: Any process that can be given a mathematical description (or a "precise enough characterization as a set of steps," or that is scientifically describable or scientifically explicable) can be simulated by a Turing machine.

As with thesis M, thesis S is trivially false if it is taken to concern all conceivable processes, and its truth-value is unknown if it is taken to concern only processes that conform to the physics of the real world. For all we presently know, a completed neuroscience may present the mind–brain as a machine that – when abstracted out from sources of inessential boundedness, such as mortality – generates functions that no Turing machine can generate.

The equivalence fallacy

Paramount among the evidence for the Church–Turing thesis properly so called is the fact that all attempts to give an exact analysis of the intuitive notion of an effective method have turned out to be *equivalent*, in the sense that each analysis has been proved to pick out the same class of functions, namely those that are computable by Turing machine. (For example, there have been analyses in terms of lambda-definability, recursivenes, register machines, Post's canonical and normal systems, combinatory definability, Markov algorithms, and Gödel's notion of reckonability.) Because of the diversity of these various analyses, their equivalence is generally considered very strong evidence for the Church–Turing thesis (although for a skeptical point of view see Kreisel 1965: 144).

However, the equivalence of these diverse analyses is sometimes taken to be evidence also for stronger theses like M and S. This is nothing more than a confusion – the *equivalence fallacy* (Copeland 2000). The analyses under discussion are of the notion of an effective method, not of the notion of a machine-generable function; the equivalence of the analyses bears only on the issue of the extent of the former notion and indicates nothing concerning the extent of the latter.

Artificial intelligence and the equivalence fallacy

Newell, discussing the possibility of artificial intelligence, argues that (what he calls) a "physical symbol system" can be organized to exhibit

general intelligence. A "physical symbol system" is a universal Turing machine, or any equivalent system, situated in the physical – as opposed to the conceptual – world. (The tape of the machine is accordingly finite; Newell specifies that the storage capacity of the tape [or equivalent] be unlimited in the practical sense of finite yet not small enough to "force concern.")

> A [physical symbol] system always contains the potential for being any other system if so instructed. Thus, a [physical symbol] system can become a generally intelligent system. (Newell 1980: 170)

Is the premise of this pro-AI argument true? A physical symbol system, being a *universal* Turing machine situated in the real world, can, if suitably instructed, simulate (or, metaphorically, become) any other physical symbol system (*modulo* some fine print concerning storage capacity). If this is what the premise means, then it is true. However, if taken literally, the premise is false, since as previously remarked, systems can be specified which no Turing machine – and so no physical symbol system – can simulate. However, if the premise is interpreted in the former manner, so that it is true, the conclusion fails to follow from the premise. Only to one who believes, as Newell does, that "the notion of machine or determinate physical mechanism" is "formalized" by the notion of a Turing machine (ibid.) will the argument appear deductively valid.

Newell's defense of his view that the universal Turing machine exhausts the possibilities of mechanism involves an example of the equivalence fallacy:

> [An] important chapter in the theory of computing . . . has shown that all attempts to . . . formulate . . . general notions of mechanism . . . lead to classes of machines that are equivalent in that they encompass in toto exactly the same set of input–output functions. In effect, there is a single large frog pond of functions no matter what species of frogs (types of machines) is used. . . . A large zoo of different formulations of maximal classes of machines is known by now – Turing machines, recursive functions, Post canonical systems, Markov algorithms . . . (Newell 1980: 150)

Newell's *a priori* argument for the claim that a physical symbol system can become generally intelligent founders in confusion.

Conclusion

Since there are problems that cannot be solved by Turing machine, there are – given the Church–Turing thesis – limits to what can be accomplished by any form of machine that works in accordance with effective methods. However, not all *possible* machines share those limits. It is an open empirical question whether there are actual deterministic physical processes that, in the long run, elude simulation by Turing machine; and, if so, whether any such processes could usefully be harnessed in some form of calculating machine. It is, furthermore, an open empirical question whether any such processes are involved in the working of the human brain.

References

Abramson, F. G. 1971. "Effective computation over the real numbers." *Twelfth Annual Symposium on Switching and Automata Theory*. Northridge, CA: Institute of Electrical and Electronics Engineers.

Boden, M. A. 1988. *Computer Models of Mind*. Cambridge: Cambridge University Press.

Boolos, G. S. and Jeffrey, R. C. 1980. *Computability and Logic*, 2nd ed. Cambridge: Cambridge University Press.

Church, A. 1936a. "An unsolvable problem of elementary number theory." *American Journal of Mathematics* 58: 345–63.

——. 1936b. "A note on the Entscheidungsproblem." *Journal of Symbolic Logic* 1: 40–1.

——. 1937. Review of Turing 1936. *Journal of Symbolic Logic* 2: 42–3.

Churchland, P. M. and Churchland, P. S. 1983. "Stalking the wild epistemic engine." *Nous* 17: 5–18.

—— and ——. 1990. "Could a machine think?" *Scientific American* 262 (Jan.): 26–31.

Copeland, B. J. 1996. "The Church–Turing Thesis." In E. Zalta, ed., *The Stanford Encyclopaedia of Philosophy*, <http://plato.stanford.edu>.

——. 1998. "Turing's O-machines, Penrose, Searle, and the Brain." *Analysis* 58: 128–38.

——. 2000. "Narrow versus wide mechanism, including a re-examination of Turing's views on the mind–machine issue." *Journal of Philosophy* 97: 5–32. Repr. in M. Scheutz, ed., *Computationalism: New Directions*. Cambridge, MA: MIT Press, 2002.

——. 2001. "Colossus and the dawning of the computer age." In M. Smith and R. Erskine, eds., *Action This Day*. London: Bantam.

——. and Proudfoot, D. 1999a. "Alan Turing's forgotten ideas in computer science." *Scientific American* 280 (April): 76–81.

—— and ——. 1999b. "The legacy of Alan Turing." *Mind* 108: 187–95.

—— and ——. 2000. "What Turing did after he invented the universal Turing machine." *Journal of Logic, Language, and Information* 9: 491–509.

—— and Sylvan, R. 1999. "Beyond the universal Turing machine." *Australasian Journal of Philosophy* 77: 46–66.

Davis, M. 1958. *Computability and Unsolvability*. New York: McGraw-Hill.

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton: Harvester.

——. 1991. *Consciousness Explained*. Boston: Little, Brown.

Deutsch, D. 1985. "Quantum theory, the Church–Turing principle and the universal quantum computer." *Proceedings of the Royal Society*, Series A, 400: 97–117.

Dreyfus, H. L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Fodor, J. A. 1981. "The mind–body problem." *Scientific American* 244 (Jan.): 124–32.

Gandy, R. 1980. "Church's thesis and principles for mechanisms." In J. Barwise, H. Keisler, and K. Kunen, eds., *The Kleene Symposium*. Amsterdam: North-Holland.

——. 1988. "The confluence of ideas in 1936." In R. Herken, ed., *The Universal Turing Machine: A Half-century Survey*. Oxford: Oxford University Press.

Geroch, R. and Hartle, J. B. 1986. "Computability and physical theories." *Foundations of Physics* 16: 533–50.

Gödel, K. 1965. "Postscriptum." In M. Davis, ed., *The Undecidable*. New York: Raven, pp. 71–3.

Gregory, R. L. 1987. *The Oxford Companion to the Mind*. Oxford: Oxford University Press.

Guttenplan, S. 1994. *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Hardy, G. H. 1929. "Mathematical proof." *Mind* 38: 1–25.

Henry, G. C. 1993. *The Mechanism and Freedom of Logic*. Lanham, MD: University Press of America.

Hilbert, D. 1902. "Mathematical problems: lecture delivered before the International Congress of Mathematicians at Paris in 1900." *Bulletin of the American Mathematical Society* 8: 437–79.

——. 1927. "Die Grundlagen der Mathematik" [The Foundations of Mathematics]. English trans. in J. van Heijenoort, ed., *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA: Harvard University Press, 1967.

Hodges, A. 1992. *Alan Turing: The Enigma*. London: Vintage.

Johnson-Laird, P. 1987. "How could consciousness arise from the computations of the brain?" In C. Blakemore and S. Greenfield, eds., *Mindwaves*. Oxford: Blackwell.

Kalmár, L. 1959. "An argument against the plausibility of Church's thesis." In A. Heyting, ed., *Constructivity in Mathematics*. Amsterdam: North-Holland.

Kleene, S. C. 1952. *Introduction to Metamathematics*. Amsterdam: North-Holland.

——. 1967. *Mathematical Logic*. New York: Wiley.

Kreisel, G. 1965. "Mathematical logic." In T. L. Saaty, ed., *Lectures on Modern Mathematics*, vol. 3. New York: Wiley.

Langton, C. R. 1989. "Artificial life." In Langton, ed., *Artificial Life*. Redwood City: Addison-Wesley.

Mendelson, E. 1964. *Introduction to Mathematical Logic*. New York: Van Nostrand.

Newell, A. 1980. "Physical symbol systems." *Cognitive Science* 4: 135–83.

Odifreddi, P. 1989. *Classical Recursion Theory*. Amsterdam: North-Holland.

Péter, R. 1959. "Rekursivität und Konstruktivität." In A. Heyting, ed., *Constructivity in Mathematics*. Amsterdam: North-Holland.

Putnam, H. 1992. *Renewing Philosophy*. Cambridge, MA: Harvard University Press.

Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

——. 1997. *The Mystery of Consciousness*. New York: New York Review of Books.

Sieg, W. 1994. "Mechanical procedures and mathematical experience." In A. George, ed.,

Mathematics and Mind. Oxford: Oxford University Press.

Sipser, M. 1997. *Introduction to the Theory of Computation*. Boston: PWS Publishing.

Smolensky, P. 1988. "On the proper treatment of connectionism." *Behavioral and Brain Sciences* 11: 1–23.

Stewart, I. 1991. "Deciding the undecidable." *Nature* 352: 664–5.

Turing, A. M. 1936. "On computable numbers, with an application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, series 2, 42 (1936–7): 230–65.

——. 1937. "Computability and λ-definability." *Journal of Symbolic Logic* 2: 156–63.

——. 1948. "Intelligent machinery." National Physical Laboratory Report. In B. Meltzer and D. Michie, eds., *Machine Intelligence 5*. Edinburgh: Edinburgh University Press, 1969. [A digital facsimile is available in The Turing Archive for the History of Computing, <http://www.AlanTuring.net/intelligent_machinery>.]

——. 1950. "Programmers' handbook for Manchester electronic computer." University of Manchester Computing Laboratory. [A digital facsimile is available in The Turing Archive for the History of Computing, <http://www.AlanTuring. net/programmers_handbook>.]

von Neumann, J. 1927. "Zur Hilbertschen Beweistheorie" [On Hilbert's proof theory], *Mathematische Zeitschrift* 26: 1–46.

Weyl, H. 1944. "David Hilbert and his Mathematical Work," *Bulletin of the American Mathematical Society* 50: 612–54.

Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology*, vol. 1. Oxford: Blackwell.

Chapter 2

Complexity

Alasdair Urquhart

1 Introduction

The theory of computational complexity is concerned with estimating the resources a computer needs to solve a problem. The basic resources are time (number of steps in a computation) and space (amount of memory used). There are problems in computer science, logic, algebra, and calculus that are solvable in principle by computers, but, in the worst case, require completely infeasible amounts of space or time, so that in practical terms they are insoluble. The goal of complexity theory is to classify problems according to their complexity, particularly problems that are important in applications such as cryptology, linear programming, and combinatorial optimization. A major result of the theory is that problems fall into strict hierarchies when categorized in accordance with their space and time requirements. The theory has been less successful in relating the two basic measures; there are major open questions about problems that are solvable using only small space, but for which the best algorithms known use exponential time.

The theory discussed in this chapter should be distinguished from another area often called "complexity theory," a loosely defined interdisciplinary stream of research that includes work on complex dynamical systems, chaos theory, artificial life, self-organized criticality, and many other subjects. Much of this research is centered in the Santa Fe Institute in New Mexico, where work on "complex systems" of various kinds is done. The confusion between the two fields arises from the fact that the word "complexity" is often used in different ways. A system or object could reasonably be described as "complex" under various conditions: if it consists of many interacting parts; if it is disordered or exhibits high entropy; if it exhibits diversity based on hierarchical structure; if it exhibits detail on many different scales, like fractal sets. Some of these meanings of "complexity" are connected with the theory of computational complexity, but some are only tangentially related. In the present chapter, we confine ourselves to the simple quantitative measures of time and space complexity of computations.

A widely accepted working hypothesis in the theoretical computer science community is that practically feasible algorithms can be identified with those whose running time can be bounded by a polynomial in the size of the input. For example, an algorithm that runs in time $10n$ for inputs with n symbols would be very efficient; this would be described as an algorithm running in linear time. A quadratic time algorithm runs in time cn^2 for some constant c; obviously such an algorithm is considerably less efficient than a linear time algorithm, but could be quite practical for inputs of reasonable size. On the other

hand, a computer procedure requiring time exponential in the size of the input very rapidly leads to infeasible running times.

To illustrate the point of the previous paragraph, consider a modern fast computer. The speed of such machines is often measured in the number of numerical operations performed per second; a commonly used standard is the number of floating-point operations per second. Suppose we have a machine that performs a million floating-point operations per second, slow by current supercomputer standards. Then an algorithm that requires n^2 such operations for an input of size n would take only a quarter of a second for an input of size 500. Even if the running time is bounded by n^3, an input of size 500 would require at most 2 minutes 5 seconds. On the other hand, an algorithm running in time 2^n could in the worst case take over 35 years for an input of size 50. The reader can easily verify with the help of a pocket calculator that this dramatic difference between polynomial and exponential growth is robust, in the sense that a thousand-fold increase in computer speed only adds 10 to the size of the largest problem instance we can solve in an hour with an exponential (2^n) time algorithm, whereas with a quadratic (n^2) time algorithm, the largest such problem increases by a factor of over 30.

The theory of computational complexity has provided rigorous proofs of the existence of computational problems for which such exponential behavior is unavoidable. This means that for such problems, there are infinitely many "difficult" instances, for which any algorithm solving the problem must take an exponentially long time. An especially interesting and important class of problems is the category of NP-complete problems, of which the satisfiability problem of propositional logic is the best-known case. These problems all take the form of asking for a solution of a certain set of constraints (formulas of propositional logic, in the case of the satisfiability problem), where a proposed solution can be quickly checked to see if it is indeed a solution, but in general there are exponentially many candidate solutions. As an example of such a problem, consider the problem of coloring a large and complicated map with only three colors so that no two countries with a common border are colored alike (see below for more details on this problem). The only known general algorithms for such problems require exponentially long run-times in the worst case, and it is widely conjectured that no polynomial time algorithms exist for them. This conjecture is usually phrased as the inequality "P ≠ NP," the central open problem in theoretical computer science, and perhaps the most important open problem in mathematical logic.

In this chapter, we begin by giving an outline of the basic definitions and results of complexity theory, including the existence of space and time hierarchies, then explain the basics of the theories of NP-completeness and parallel computation. The chapter concludes with some brief reflections on the relevance of complexity theory to questions in the philosophy of computing.

2 Time and Space in Computation

The theory of complexity analyzes the computational resources necessary to solve a problem. The most important of these resources are time (number of steps in a computation) and space (storage capacity of the computer). This chapter is mainly concerned with the complexity of decision problems having infinitely many instances. There is another approach to complexity applicable to individual objects, in which the complexity of an object is measured by the size of the shortest program that produces a description of it. This is the Kolmogorov complexity of the object; Li and Vitányi (1997) give a readable and detailed introduction to this subject in their textbook.

The model for computation chosen here is the Turing machine, as defined in the preceding chapter. The time for a computation is the number of steps taken before the machine halts; the space is the number of cells of the tape visited by the reading head during the computation. Several other models of sequential computation have been proposed. The time and space complexity of a problem clearly depend on the machine model adopted. However, the basic concepts of complexity theory defined here are

robust in the sense that they are the same for any reasonable model of sequential computation.

Let Σ be a finite alphabet, and Σ^* the set of all finite strings in this alphabet. A subset of Σ^* is said to be a problem (often called a "language"), and a string in Σ^* an instance of the problem. The size $|s|$ of an instance s is its length, i.e. the number of occurrences of symbols in it. A function f defined on Σ^* and having strings as its values is computed by a Turing machine M if for any string s in Σ^*, if M is started with s on its tape, then it halts with $f(s)$ on its tape. A problem L is solvable (decidable) if there is a Turing machine that computes the characteristic function of L (the function f such that $f(s) = 1$ if s is in L and $f(s) = 0$ otherwise). For example, the satisfiability problem of determining whether a formula of propositional logic is satisfiable or not is solvable by the familiar method of truth-tables.

Solvable problems can be classified according to the time and space required for their solution. If f is a computable function, then we say that f is computable in time $T(n)$ if there is a Turing machine computing f that for any input s halts with output $f(s)$ after $O(T(|s|))$ steps (that is, there is a constant c such that M halts in at most $c \cdot T(|s|)$ steps). Similarly, f is computable in space $T(n)$ if there is a machine M computing f so that for any input s, M halts after visiting $O(T(|s|))$ squares on its tape. A problem L is solvable in time $T(n)$ if the characteristic function of L is computable in time $T(n)$; L is solvable in space $S(n)$ if the characteristic function of L is computable in space $S(n)$. For example, the truth-table method shows that the satisfiability problem can be solved in time 2^n and space n (we need only enough tape space to evaluate the truth-table one row at a time).

As an illustration of these rather abstract definitions, let us consider a concrete problem. Suppose that our alphabet contains only two symbols, so that $\Sigma = \{a; b\}$, and Σ^* is the set of all finite strings consisting of a's and b's. The palindrome problem *PAL* is defined by letting the instances in *PAL* be all those strings in Σ^* that read the same forward as backwards; for example, *aba* and *bbb* are both palindromes, but *ab* and *bba* are not. This problem can be solved in time n^2 by a simple strategy that involves checking the first against the last symbol, deleting these two symbols, and repeating this step until either the empty string (with no symbols at all) or a string consisting of exactly one symbol is reached. (This is an instructive exercise in Turing machine programming.) In fact, it is not possible to do much better than this simple algorithm. Any algorithm for a single-tape Turing machine requires cn^2 steps to solve *PAL* for some $c > 0$; for an elegant proof of this fact using the "incompressibility method" see Li and Vitányi (1997: ch. 6).

Other natural examples of computational problems arise in the area of games. For example, given a chess position, consider the problem: "Is this a winning position for White?"; that is to say, does White have a plan that forces checkmate no matter how Black plays? In this case, there is a simple but crude algorithm to answer any such question – simply compile a database of all possible board positions, then classify them as winning, losing, or drawing for White by considering all possible continuations. Such databases have been compiled for the case of endgames with only a few pieces (for example, queen versus rook endgames). Can we do better than this brute-force approach? There are reasons to think not. The results of Fraenkel and Lichtenstein described below show that computing a perfect strategy for a generalization of chess on an n by n board requires time exponential in n.

One of the most significant complexity classes is the class P of problems solvable in polynomial time. A function f is polynomial-time computable if there exists a polynomial p for which f is computable in time $p(n)$. A problem is solvable in polynomial time if its characteristic function is polynomial-time computable. The importance of the class rests on the widely accepted working hypothesis that the class of practically feasible algorithms can be identified with those algorithms that operate in polynomial time. Similarly, the class PSPACE contains those problems solvable in polynomial space. The class EXPTIME consists of the problems solvable in exponential time; a problem is solvable in exponential time if there is a k for which it is solvable in time 2^{n^k}. The class EXPSPACE contains those problems solvable in exponential space. The satisfiability problem is in EXPTIME; whether it is in P is a major open problem.

3 Hierarchies and Reducibility

A fundamental early result of complexity theory is the existence of strict hierarchies among problems. So, for example, we can prove that there are problems that can be solved in time n^2, but not in time n, and similar theorems hold for space bounds on algorithms. To state this result in its most general form, we introduce the concept of a space constructible function. A function $S(n)$ is said to be space constructible if there is a Turing machine M that is $S(n)$ space bounded, and for each n there is an input of length n on which M actually uses $S(n)$ tape cells. All "reasonable" functions such as n^2, n^3, and 2^n are space constructible. The space hierarchy theorem, proved by Hartmanis, Lewis and Stearns in 1965, says that if $S_1(n)$ and $S_2(n)$ are space constructible functions, and S_2 grows faster than S_1 asymptotically, so that

$$\liminf_{n\to\infty} \frac{S_1(n)}{S_2(n)} = 0,$$

then there exists a problem solvable in space $S_2(n)$, but not in space $S_1(n)$. A similar hierarchy theorem holds for complexity classes defined by time bounds. The hard problems constructed in the proofs of the hierarchy theorems are produced by diagonalizing over classes of machines, and so are not directly relevant to problems arising in practice. However, we can prove lower bounds on the complexity of such problems by using the technique of efficient reduction. We wish to formalize the notion that one problem can be reduced to another in the sense that if we had an efficient algorithm for the second problem, then we would have an efficient algorithm for the first. Let L_1 and L_2 be problems expressed in alphabets Σ_1 and Σ_2. L_1 is said to be polynomial-time reducible to L_2 (briefly, reducible to L_2) if there is a polynomial-time computable function f from Σ_1^* to Σ_2^* such that for any string s in Σ_1^*, s is in L_1 if and only if $f(s)$ is in L_2. Other notions of reducibility can be defined by varying the class of functions f that implement the reduction. The importance of the concept lies in the fact that if we have an efficient algorithm solving the problem L_2, then we can use the function f to produce an efficient algorithm for L_1. Conversely, if there is no efficient algorithm for L_1, then there cannot be an efficient algorithm for L_2. Notice that the class P is closed under polynomial-time reductions since if L_1 is reducible to L_2, and L_2 is in P, then L_1 is also in P.

If C is a complexity class, and L is a problem in C so that any problem in C is reducible to L, then L is said to be C-complete. Such problems are the hardest problems in C; if any problem in C is computationally intractable, then a C-complete problem is intractable. The technique of reducing one problem to another is very flexible, and has been used to show a large variety of problems in computer science, combinatorics, algebra, and combinatorial game theory intractable. We now provide some examples of such problems.

The time hierarchy theorem implies that there are problems in EXPTIME that require exponential time for their solution, no matter what algorithm is employed. The reduction method then allows us to draw the same conclusion for other problems. For example, let us define generalized chess to be a game with rules similar to standard chess, but played on an $n \times n$ board, rather than an 8×8 board. Fraenkel and Lichtenstein (1981) used the reduction technique to show that generalized chess is EXPTIME-complete, and hence computationally intractable.

EXPSPACE-complete problems are also computationally intractable. An example of a problem of this type in classical algebra is provided by the word problem for commutative semigroups. Here the problem is given in the form of a finite set of equations formed from a set of constants using a single binary operation that is assumed to be associative and commutative, together with a fixed equation $s = t$. The problem is to determine whether $s = t$ is deducible from the set of equations, assuming the usual rules for equality. Mayr and Meyer in 1981 showed this problem to be EXPSPACE-complete, so that any algorithm solving this problem must use an exponential amount of space on infinitely many inputs.

Logic also provides a fertile source of examples of intractable problems. Although the decision problem for true sentences of number theory is unsolvable, if we restrict ourselves to sentences

that involve only the constants 0 and 1, together with identity and the addition symbol, then there is an algorithm to determine whether such a sentence is true or false, a result proved by Presburger in 1930. However, in 1973 Rabin and Fischer showed that the inherent complexity of this problem is doubly exponential. This means that for any machine solving this problem, there is a constant $c > 0$ so that for infinitely many sentences the machine takes at least $2^{2^{cn}}$ steps to determine whether it is true or not.

If we add quantification over finite sets, then we can prove even more powerful lower bounds. The weak monadic second-order theory of one successor (WS1S) is formulated in a second-order language with equality, the constant 0 and a successor function. In the intended interpretation for this theory, the second-order quantifiers range over finite sets of non-negative integers. The decision problem for this theory was proved to be solvable by Büchi in 1960, but its inherent complexity is very high. Albert Meyer showed in 1972 that an algorithm deciding this theory must use for infinitely many inputs of length n an amount of space that is bounded from below by an iterated exponential function, where the stack contains at least dn iterations, for a fixed $d > 0$.

The conclusion of the previous paragraph could be challenged by pointing out that Meyer's lower bound is an asymptotic result that does not rule out a practical decision procedure for sentences of practically feasible size. However, a further result shows that astronomical lower bounds can be proved for WS1S even if we restrict the length of sentences. A Boolean network or circuit is an acyclic directed graph in which the nodes are labeled with logical operators such as AND, OR, NOT etc. Such a network with designated input and output nodes computes a Boolean function in an obvious way. Meyer and Stockmeyer showed that any such network that decides the truth of all sentences of WS1S of length 616 or less must contain at least 10^{123} nodes. Even if the nodes were the size of a proton and connected by infinitely thin wires, the network would densely fill the known universe.

Inherently intractable problems also exist in the area of nonclassical propositional logics. The area of substructural logics, such as linear logic and relevance logics, provides us with several such examples. The implication-conjunction fragment of the logic **R** of relevant implication was proved decidable by Saul Kripke in 1959 using a sophisticated combinatorial lemma. The author of the present chapter showed (Urquhart 1999) that this propositional logic has no primitive recursive decision procedure, so that Kripke's intricate method is essentially optimal. This is perhaps the most complex decidable nonclassical logic known.

4 NP-completeness and Beyond

A very common type of computational problem consists in searching for a solution to a fixed set of conditions, where it is easy to check whether a proposed solution really is one. Such solutions may be scattered through a very large set, so that in the worst case we may be reduced to doing an exhaustive search through an exponentially large set of possibilities. Many problems of practical as well as theoretical interest can be described in this general setting. The theory of NP-completeness derives its central importance in computer science from its success in providing a flexible theoretical framework for this type of problem.

A problem L belongs to the class NP if there is a polynomial p and a polynomial-time computable relation R so that a string x is in L if and only if there is a string y so that the length of y is bounded by $p(|x|)$, and $R(x, y)$ holds. The idea behind the definition is that we think of y as a succinct proof (or 'certificate') that x is in P, where we insist that we can check efficiently that an alleged proof really is a proof.

Here are a few examples to illustrate this definition. Consider the problem of determining whether an integer in decimal notation is nonprime (that is to say, the strings in the problem are the decimal representations of numbers that are not prime). Then a proof that a number x is not prime consists of a pair of numbers $y, z > 1$ so that $yz = x$. The satisfiability problem is also easily seen to be in NP. Here a positive instance of the problem consists of a satisfiable formula F of propositional logic; the proof that F is satisfiable is simply a line of a truth-table. It is obvious that we can check very quickly if a formula

is satisfied by an assignment; on the other hand, the best current algorithms for satisfiability in the worst case are forced to check exponentially many possibilities, thus being not much different from the crude brute-force method of trying all possible lines in the truth-table for a given formula.

The satisfiability problem occupies a central place in theoretical computer science as the best known NP-complete problem. Any problem in NP can be reduced efficiently to the satisfiability problem. This reflects the fact that the language of propositional logic forms a kind of universal language for problems of this type. Given a problem in NP, it is usually a routine exercise to see how to translate the problem into a set of conditions in propositional logic so that the problem has a solution if and only if the set of conditions is satisfiable. For example, consider the problem of coloring a map in the plane with three colors. Here the problem takes the form of a map, and a set of three colors, say red, white, and blue, so that adjacent countries are colored differently. We can formalize this problem by introducing a set of constants to stand for the countries in the map, and variables Rx, Wx, Bx to stand for "Country x is colored red (white, blue)." The reader should check that given a map, we can quickly write down a corresponding set of conditions in propositional logic that formalizes the statement that the map can be properly colored with the three colors.

Cook's famous theorem of 1971 showing that satisfiability is NP-complete was quickly followed by proofs that many other well-known computational problems fall into this class. Since then, thousands of significant problems have been proved NP-complete; for a partial list, see the book by Garey and Johnson (1979). The ubiquity of NP-completeness in the theory of combinatorial problems means that a proof of P = NP (that is to say, a proof that there is a polynomial-time algorithm for satisfiability) would have dramatic consequences. It would mean that feasible solutions would exist for hundreds of problems that are currently intractable. For example, the RSA cryptosystem, widely used for commercial transactions on the internet, would immediately be vulnerable to computer attack, since the security of the system rests on the assumed intractability of the problem of factoring a number that is the product of two large prime numbers. The same remarks apply to other cryptosystems, with the exception of the theoretically invulnerable one-time pad system. The fact that no such feasible algorithm has been found for any of these problems is one of the main reasons for the widespread belief in the conjecture that P ≠ NP.

The lower bounds described in the preceding section were all proved by the diagonal method. That is to say, the method in each case was an adaptation of the technique originally employed by Cantor to prove the set of real numbers uncountable, and subsequently adapted by Church and Turing to prove the decision problem for predicate logic unsolvable. There are reasons to think that this method will not succeed in resolving the problem of whether or not P = NP. To explain these reasons, we need to introduce the concept of a Turing machine with an oracle. The concept of a Turing machine explicates the notion of computability in an absolute sense. Similarly, the concept of an oracle machine explicates the general notion of what it means for a problem to be solvable relative to another problem (the definition of reducibility above is a special case of this general notion). If A is a set of strings then a Turing machine with oracle A is defined to be a Turing machine with three special states $q_?$, q_y, and q_n. The query state $q_?$ is used to ask "Is the string of nonblank symbols to the right of the reading head in A?" The answer is supplied by having the machine change on the next move to one of the two states q_y or q_n, depending on whether the answer is yes or no. Time and space of a computation by an oracle machine are computed just as for an ordinary Turing machine, counting the time taken for the answer to the oracle query as one step (the oracle answers any query instantaneously).

We can imagine an oracle machine as representing a situation where we have access to a "black box" that instantaneously answers questions belonging to a type for which we have no algorithm, or for which the only known algorithm is very inefficient. For example, suppose that the oracle (black box) can answer all queries of the form: "Do all integers n satisfy the property $P(n)$?," where P is a decidable property of integers. Then the black box exhibits a

kind of limited omniscience that would enable us to answer instantaneously many open problems of current mathematics such as Goldbach's conjecture or the Riemann hypothesis. In spite of this, it is possible to show that there are problems that such a miraculous machine cannot answer; classical recursion theory (computability theory) is largely taken up with such problems.

If A is any set of strings, then by imitating the definitions of the complexity classes above, but substituting "Turing machine with oracle A" everywhere for "Turing machine" we can define relativized complexity classes P(A), NP(A), and so on. Baker, Gill, and Solovay proved in 1975 that there is a decidable oracle A for which P(A) = NP(A), and a decidable oracle B for which P(B) ≠ NP(B). The significance of this theorem lies in the fact that known techniques of diagonalization, such as are used in computability theory, continue to work in the presence of oracles. Thus it provides evidence that standard diagonal techniques are inadequate to settle such questions as "P = NP?"

The literature of theoretical computer science contains many complexity classes beyond the few discussed here; for details, the reader should consult the collection of survey articles in Van Leeuwen (1990). We conclude this section with a brief description of an important complexity class that, like the classes P and NP, has strong connections with logic. The class PSPACE consists of those problems solvable using a polynomial amount of space. It is not hard to see that this class contains the class NP, since we require only a small amount of space to do an exhaustive search through the space of all possible strings that are candidates for certificates showing that a string is a positive instance of an NP-complete problem. This class of problems bears the same relationship to the quantified propositional calculus as the class NP to the ordinary propositional calculus. In the quantified propositional calculus, we add to ordinary propositional logic quantifiers ranging over propositions. Thus, for example, the formula $\exists p \forall q (p \rightarrow q)$ is a logical truth in this language. The valid (logically true) formulas of quantified propositional logic constitute a PSPACE-complete set (that is, the problem of determining the validity of formulas in the quantified language is PSPACE-complete).

The family of algorithms operating in polynomial space appears to be a much more extensive class than the family of algorithms operating in polynomial time. However, we are unable on the basis of current knowledge to refute the equality P = PSPACE. This illustrates the point mentioned in the introduction, that in contrast to the detailed hierarchy theorems known for time and space separately, the problem of relating time and space requirements for computations remains largely unsolved.

5 Parallel Computation

The computational model discussed in the preceding sections was that of serial or sequential computation, where the machine is limited to a bounded number of actions at each step, for example, writing a symbol, moving left or right, and changing internal state in the case of the Turing model. However, there is considerable current interest, both theoretical and practical, in parallel models of computation. Parallel computation is attractive in applications such as searching large databases, and also is of interest in modeling brain function (since the brain seems, to a first approximation, to be some kind of parallel computer). In this section, we provide a brief discussion of the complexity of parallel computation.

In the case of parallel computation, various models have been proposed, and there is no universal agreement on the best. These include models such as the PRAM (parallel random access machine), where a large number of simple processors with limited memory have joint access to a large shared memory; various conventions on read–write conflicts can be adopted. For more details on these models, the readers should consult the articles of van Emde Boas, Karp, and Ramachandran in Van Leeuwen (1990). We shall not discuss these models further here, but instead describe the area of non-uniform complexity.

The Turing model has the property that a single machine operates on inputs of arbitrary length. An alternative approach to measuring the complexity of computations is to limit ourselves to functions of a fixed input and output

size – Boolean functions, in the case of decision problems – and then estimate the minimum size of the circuitry needed to provide a "hard-wired" version of the function.

We define a circuit as a finite, labeled, directed graph with no directed cycles. The nodes with no arrows pointing in are input nodes, while the nodes with no arrows pointing out are output nodes. The internal nodes are considered as logic gates, and labeled with appropriate Boolean functions. For example, a circuit could be built from AND gates with two inputs and one output, and NOT gates with one input and one output. The important complexity measures for a circuit are its depth (length of the shortest path from an input node to an output node) and its size (number of nodes in the circuit).

We can now define parallel complexity classes using the circuit model. Perhaps the most important of these is the class of problems with polynomial-size circuits, abbreviated as P/poly. Given a problem L, we can encode the strings of L in binary notation; let us refer to this encoded problem as L_b. Then L is said to have polynomial-size circuits if there is a polynomial p so that for every n there is a Boolean circuit C with size bounded by $p(n)$ so that C gives the output 1 exactly for those binary strings of length n that belong to L_b, that is, exactly those strings of length n that represent positive instances of the problem L.

This is a much more powerful model of computation than the standard Turing model; it is non-uniform, since we allow a different circuit for each input length n. In particular, it is not hard to see that in this model, an unsolvable problem can have polynomial-size circuits.

The connection between the circuit model and the Turing model can be made more precise by considering the oracle machines defined earlier. Given a fixed circuit, we can easily program a Turing machine to simulate its behavior, simply by encoding the circuit as a big look-up table (we discuss the philosophical import of this observation below). Hence, if a problem L has polynomial-size circuits, we can program an oracle machine that, relative to the oracle set C representing the encoding of the family of circuits solving L, solves the problem. The machine can be considered as a machine that takes a "polynomially bounded amount of advice"; conversely, any problem solved by such a machine has polynomial-size circuits.

The description of P/poly in the preceding paragraph should make it clear that we are dealing with an extremely powerful class of procedures, since they have the ability to answer arbitrarily complex questions about finite configurations in the time it takes to write down the question (and so should be considered as "algorithms" only in an extended sense). Nevertheless, it is widely conjectured that the satisfiability problem does not have polynomial-size circuits. Current proof techniques in the theory of Boolean circuits seem to be inadequate for resolving this challenging conjecture.

6 Complexity and Philosophy

Philosophical treatments of the concept of computation often ignore issues relating to complexity. However, we shall argue in this section that such questions are directly relevant to some frequently discussed problems. Since Turing's famous article of 1950, it has been common to replace the question "Can machines think?" – which Turing thought too meaningless to discuss – with the question "Can digital computers successfully play the imitation game?" Turing made the following optimistic prediction:

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. (Turing 1950)

It is clear that Turing was thinking of computers as real physical devices. However, let us suppose that for the moment we think of computers as idealized mathematical machines, and (as is common in the mathematical context of computability theory) ignore all questions of resources, efficiency, and so forth. Then it is a mathematical triviality that the answer to Turing's

question is affirmative. Let us recall the basic situation for the imitation game. An interrogator communicates by teletype with two participants, one a human being, the other a digital computer. The task of the interrogator is to determine by skillful questioning which of the two is the human being. For a computer to succeed at the imitation game means that it can succeed in fooling the interrogator in a substantial number of cases, if the game is played repeatedly.

Turing envisages a limit of five minutes of interrogation, but for our present purposes, let us suppose that we simply limit the number of symbols exchanged between the participants in the game to some reasonably large number (bearing in mind that all the participants have to type at human speeds, otherwise the computer could be spotted immediately). It is now easy to see that there is indeed a machine (in the mathematical sense) that can play this game with perfect success (i.e. a skilled interrogator cannot improve on random guessing in the game). Consider all sequences of symbols representing a possible sequence of questions and answers in the imitation game. Of these, some will be bad, in the sense that they will easily reveal to the interrogator the identity of the computer, while others are good (we can imagine these to be the sort of responses produced when the computer is replaced by a human). Now provide the computer with the set of all good sequences as a gigantic look-up table, and program the computer to answer in accordance with this table. By definition, the computer must succeed perfectly at the game.

Of course, the "machine" described in the previous paragraph is a pure mathematical abstraction, but it suffices to illustrate the fact that in philosophical, as opposed to mathematical, contexts, the purely abstract definition of a machine is not appropriate. Similar remarks apply in the case of the distinction between serial and parallel computation.

It is currently fashionable to think of cognitive processes as modeled by neural networks composed of simple elements (typically threshold gates of some kind), joined together in some random fashion, and then "trained" on some family of inputs, the "learning" process consisting of altering the strength of connections between gates. This model is sometimes described in the cognitive science literature as "parallel distributed processing" or "PDP" for short. If we take into account speed of processing, then such models may indeed provide more accurate simulations of processes in real brains, since neurophysiology indicates that mammalian brains are made out of relatively slow elements (neurons) joined together in a highly connected network. On the other hand, there is nothing new here as compared with the classical serial model of computation, if we ignore limitations of time and space. Nevertheless, some of the literature in cognitive science argues otherwise.

In their debate of 1990 with John Searle, Paul and Patricia Churchland largely agree with the conclusions of Searle's critique of classical AI (based on a serial model of computation), for which Searle argues on the grounds of his "Chinese room" thought experiment, but disagree with the conclusions of his "Chinese gym" thought experiment designed to refute the claims of parallel processors to represent the mind. The Churchlands first point out the implausibility of the simulation (involving huge numbers of people passing messages in an enormous network), but then continue:

> On the other hand, if such a system were to be assembled on a suitably cosmic scale, with all its pathways faithfully modeled on the human case, we might then have a large, slow, oddly made but still functional brain on our hands. In that case the default assumption is surely that, given proper inputs, it would think, not that it couldn't. (Churchland & Churchland 1990)

This imaginary cosmic network is a finite object, working according to a fixed algorithm (as embodied in its pattern of connections). It follows that it can be simulated by a serial computer (in fact, all of the early research on "neural nets" was carried out by writing simulation programs on computers of conventional design). Of course, there will be a loss of speed, but the Churchlands explicitly rule out speed of operation as a relevant variable. It's difficult to see, though, why the serial computer doing the simulation of the neural network should be ruled out as a "functional brain."

Let us expand a little more on the implications of this analysis. The basic fact that serial machines can simulate parallel machines (a point emphasized by Searle himself) should not be considered as an argument for or against either the Chinese-room argument or the Chinese-gym argument, both of which involve obscure questions concerning the presence or absence of "mental contents" or "semantic contents." Rather, it points to the difficulties of a position that rejects a serial model of computation for the mind, but accepts a parallel model, while ignoring questions of complexity and efficiency.

Since we are not limited by technological feasibility, let us imagine a huge, super-fast serial computer that simulates the Churchlands' cosmic network. Furthermore, to make the whole thing more dramatic, let's imagine that this marvelous machine is wired up to a gigantic cosmic network with flashing lights showing the progress of the computation, working so fast that we can't tell the difference between a real cosmic network and the big display. Is this a "functional brain" or not? It's hard to know what the criteria are for having a "functional brain on our hands," but without considering questions of computational complexity, it is difficult to see how we can reject serial candidates for "functional brains." For a more detailed discussion of the "Chinese room" argument from a complexity-theoretic perspective, the reader should consult Parberry 1994. (See also Chapter 9, The Philosophy of AI and Its Critique.)

Current work in the philosophy of mind manifests a fascination with far-fetched thought experiments, involving humanoid creatures magically created out of swamp matter, zombies, and similar imaginary entities. Philosophical discussion on the foundations of cognitive science also frequently revolves around implausible thought experiments like Searle's "Chinese room" argument. The point of the simple observations above is that unless computational resources are considered, arguments based on such imaginary experiments may appear quite powerful. On the other hand, by taking such resources into account, we can distinguish between objects that exist in the purely mathematical sense (such as the Turing machine that succeeds at the imitation game), and devices that are physically constructible.

References

Churchland, P. M. and Churchland, P. S. 1990. "Could a machine think?" *Scientific American* 262: 32–7. [A spirited reply to John Searle's article in the same issue describing his "Chinese room" thought experiment.]

Fraenkel, A. S. and Lichtenstein, D. 1981. "Computing a perfect strategy for $n \times n$ chess requires time exponential in n." *Journal of Combinatorial Theory Series A* 31: 199–214.

Garey, M. R. and Johnson D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-completeness.* San Francisco: Freeman. [This well-known text contains a very readable introduction to the theory of NP-completeness, as well as an extensive list of NP-complete problems from many areas.]

Li, M. and Vitányi, P. 1997. *An Introduction to Kolmogorov Complexity and its Applications.* New York: Springer-Verlag. [The basic textbook on this fascinating theory; it contains detailed technical developments as well as more discursive material on the general notion of complexity.]

Parberry, I. 1994. *Circuit Complexity and Neural Networks.* Cambridge, MA: MIT Press. [The first chapter is an excellent nontechnical discussion of the Chinese-room thought experiment from a complexity-theoretic point of view. The remainder of the book is a more technical, but accessible, discussion of the problem of scaling in neural network theory.]

Turing, A. 1950. "Computing machinery and intelligence." *Mind* 59: 433–60. [The classic article in which the great computer pioneer presented the "TuringTest" for machine intelligence.]

Urquhart, A. 1999. "The complexity of decision procedures in relevance logic II." *Journal of Symbolic Logic* 64: 1774–1802. [This article gives a detailed proof that the propositional logic of relevant implication, restricted to conjunction and implication, has enormous intrinsic complexity.]

Van Leeuwen, J., ed. 1990. *Handbook of Theoretical Computer Science*, Volume A: *Algorithms and Complexity.* Amsterdam: Elsevier. [A collection of detailed survey articles by leading researchers covering many topics, including parallel complexity and cryptology.]

Chapter 3

System: An Introduction to Systems Science

Klaus Mainzer

Introduction

Dynamical systems, with their astonishing variety of forms and functions, have always fascinated scientists and philosophers. Today, structures and laws in nature and society are explained by the dynamics of complex systems, from atomic and molecular systems in physics and chemistry to cellular organisms and ecological systems in biology, from neural and cognitive systems in brain research and cognitive science to societies and market systems in sociology and economics. In these cases, complexity refers to the variety and dynamics of interacting elements causing the emergence of atomic and molecular structures, cellular and neural patterns, or social and economic order (on computational complexity see Chapter 1, COMPUTATION). Computational systems can simulate the self-organization of complex dynamical systems. In these cases, complexity is a measure of computational degrees for predictability, depending on the information flow in the dynamical systems. The philosophy of modern systems science aims to explain the information and computational dynamics of complex systems in nature and society.

The first section of this chapter defines the basic concept of a dynamical system. The dynamics of systems is measured by time series and modeled in phase spaces, which are introduced in section 2. Phase spaces are necessary to recognize attractors of a system, such as chaos. In the case of chaos, severe restrictions on long-term predictions and systems control must be taken into account. But, in practice, there are only finitely many measurements and observations of a time series. So, in section 3, time-series analysis is introduced in order to reconstruct phase spaces and attractors of behavior. Section 4 presents examples of complex systems in nature and society. From a philosophical point of view, dynamical systems in nature and society can be considered as information and computational systems. This deep insight of modern systems science is discussed in the last section.

1 Basic Concepts of Systems Science

A *dynamical system* is characterized by its *elements* and the *time-depending* development of their *states*. In the simple case of a falling stone,

one may consider for example only the acceleration of a single element. In a planetary system, the states of planets are also determined by their position and momentum. The states can also refer to moving molecules in a gas, the excitation of neurons in a neural net, nutrition of populations in an ecological system, or products in a market system. The *dynamics* of a system, that is, the change of system states depending on time, is mathematically described by *differential equations*. For *deterministic processes* (e.g., motions in a planetary system), each future state is uniquely determined by the present state. A *conservative* (Hamiltonian) *system*, e.g. an ideal pendulum, is determined by the reversibility of time direction and conservation of energy. Conservative systems are closed and have no energetic dissipation with their environment. Thus, conservative systems in the strict sense exist only as approximations like, e.g., an ideal Thermos bottle. In our everyday world, we mainly observe *dissipative systems* with a distinct time direction. Dissipative systems, e.g., a real pendulum with friction, are irreversible.

In classical physics, the dynamics of a system is analyzed as a *continuous process*. In a famous quotation, Leibniz assumed that nature does not jump (*natura non facit saltus*). However, continuity is only a mathematical idealization. Actually, a scientist deals with single observations or measurements at discrete time points that are chosen equidistant or defined by other measurement devices. In discrete processes, there are finite differences between the measured states, no infinitely small differences between the measured states, and no infinitely small differences (differentials) that are assumed in a continuous process. Thus, discrete processes are mathematically described by *difference equations*.

Random events (e.g., Brownian motion in a fluid, mutation in evolution, innovations in economy) are represented by additional *fluctuation terms*. *Classical stochastic processes*, e.g. the billions of unknown molecular states in a fluid, are defined by time-depending differential equations with distribution functions of probabilistic states. In *quantum systems* of elementary particles, the dynamics of quantum states is defined by Schrödinger's equation with observables (e.g., position and momentum of a particle) depending on *Heisenberg's principle of uncertainty*. The latter principle allows only *probabilistic forecasts* of future states.

2 Dynamical Systems, Chaos, and Other Attractors

During the centuries of classical physics, the universe was considered as a deterministic and conservative system. The astronomer and mathematician Pierre-Simon Laplace (1814), for example, assumed the total computability and predictability of nature if all natural laws and initial states of celestial bodies are well known. The *Laplacean spirit* well expressed philosophers' faith in determinism and computability of the world during the eighteenth and nineteenth centuries.

Laplace was right about *linear* and *conservative dynamical systems*. A simple example is a so-called harmonic oscillator, like a mass attached to a spring oscillating regularly without friction. Let us consider this example in more detail. It will help us to introduce the basic notions of time series, phase space, and trajectory, essential to understand the structure and development of dynamical systems. In general, *a linear relation* means that the rate of change in a system is proportional to its cause: small changes cause small effects, while large changes cause large effects. In the example of a harmonic oscillator, a small compression of a spring causes a small oscillation of the position of a mass, while a large compression causes a large oscillation, following Hooke's law. Changes of a dynamical system can be modeled in one dimension by changing values of a time-depending quantity along the time axis (*time series*). In figure 3.1a, the position $x(t)$ of a mass attached to a spring is oscillating in regular cycles along the time axis t. $x(t)$ is the solution of a linear equation, according to Hooke's law. Mathematically, linear equations are completely computable. This is the deeper reason for Laplace's philosophical assumption to be right for linear and conservative systems.

In systems theory, the *complete information* about a dynamical system at a certain time is

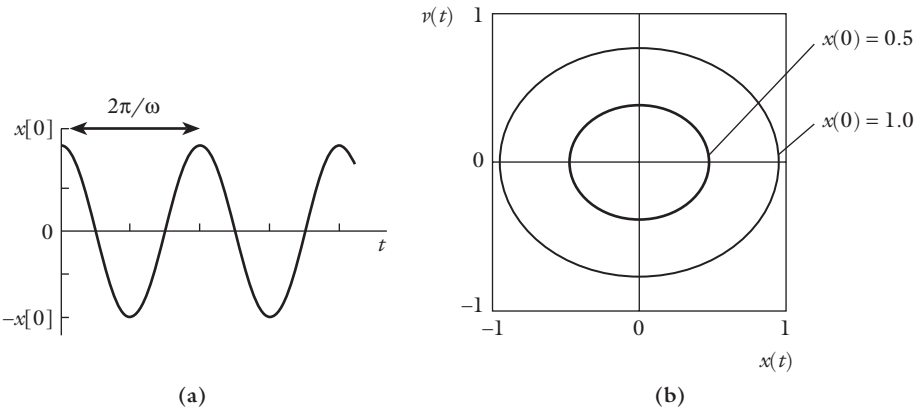

Figure 3.1a: A solution $x(t)$ of a linear equation as time series (Kaplan 1995: 211)
Figure 3.1b: Trajectories of two solutions of a linear equation in a 2-dimensional phase space (Kaplan 1995: 212)

determined by its *state* at that time. In the example of a harmonic oscillator, the state of the system is defined by the position $x(t)$ and the velocity $v(t)$ of the oscillating mass at time t. Thus, the state of the system is completely determined by a pair of two quantities that can be represented geometrically by a point in a 2-dimensional phase space, with a coordinate of position and a coordinate of velocity. The dynamics of a system refers to the time-depending development of its states. Thus, the dynamics of a system is illustrated by an orbit of points (*trajectory*) in a phase space corresponding to the time-depending development of its states. In the case of an harmonic oscillator, the trajectories are closed ellipses around a point of stability (figure 3.1b), corresponding to the periodic cycles of time series, oscillating along the time axis (figure 3.1a). Obviously, the regular behavior of a linear and conservative system corresponds to a regular and stable pattern of orbits. So, the past, present, and future of the system are completely known.

In general, the state of a system is determined by more than two quantities. This means that higher dimensional phase space is required. From a methodological point of view, time series and phase spaces are important instruments to study systems dynamics. The *state space of a system* contains the *complete information of its past, present and future behavior*. The dynamics of real systems in nature and society is, of course, more complex, depending on more quantities, with patterns of behavior that are not as regular as in the simple case of a harmonic oscillator. It is a main insight of modern systems theory that the behavior of a dynamic system can only be recognized if the corresponding state space can be reconstructed.

At the end of the nineteenth century, Henri Poincaré (1892–3) discovered that celestial mechanics is not a completely computable clockwork, even if it is considered as a deterministic and conservative system. The mutual gravitational interactions of more than two celestial bodies ("Many-bodies-problem") correspond to *nonlinear* and *non-integrable equations* with *instabilities* and *irregularities*. According to the Laplacean view, similar causes effectively determine similar effects. Thus, in the phase space, trajectories that start close to each other also remain close to each other during time evolution. Dynamical systems with *deterministic chaos* exhibit an exponential dependence on initial conditions for bounded orbits: the separation of trajectories with close initial states increases exponentially (figure 3.2).

Tiny deviations of initial data lead to exponentially increasing computational efforts to analyze future data, *limiting long-term predictions*, although the dynamics is in principle uniquely determined. This is known as the "*butterfly effect*": initial, small, and local causes soon lead to unpredictable, large and global effects (see figure

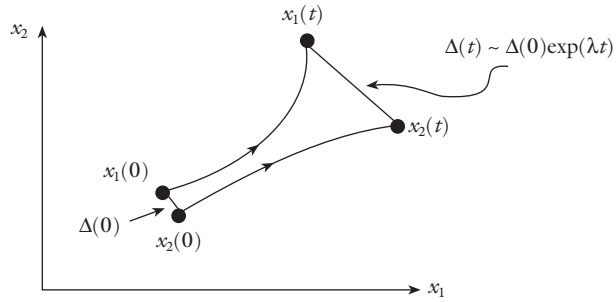

Figure 3.2: Exponential dependence of two trajectories $x_1(t)$ and $x_2(t)$ on initial conditions

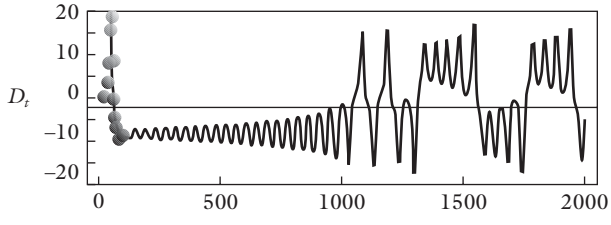

Figure 3.3: A measured signal D_t from a chaotic system (time series)

3.3). According to the famous *KAM Theorem* of A. N. Kolmogorov (1954), V. I. Arnold (1963), and J. K. Moser (1967), trajectories in the phase space of classical mechanics are neither completely regular, nor completely irregular, but depend sensitively on the chosen initial conditions.

Dynamical systems can be *classified* on the basis of the effects of the dynamics on a region of the phase space. A *conservative system* is defined by the fact that, during time evolution, the volume of a region remains *constant*, although its shape may be transformed. In a *dissipative system*, dynamics causes a *volume contraction*. An *attractor* is a region of a phase space into which all trajectories departing from an adjacent region, the so-called *basin of attraction*, tend to converge. There are different kinds of attractors. *Fixed points* form the simplest class of attractors. In this case, all trajectories of adjacent regions converge to a point. An example is a dissipative harmonic oscillator with friction: the oscillating system is gradually slowed down by frictional forces and finally comes to a rest in an equilibrium point.

Conservative harmonic oscillators without friction belong to the second class of attractors with *limit cycles*, which can be classified as being periodic or quasi-periodic. A *periodic orbit* is a closed trajectory into which all trajectories departing from an adjacent region converge. For a simple dynamical system with only two degrees of freedom and continuous time, the only possible attractors are fixed points or periodic limit cycles. An example is a Van der Pol oscillator modeling a simple vacuum-tube oscillator circuit.

In continuous systems with a phase space of dimension $n > 2$, more complex attractors are possible. Dynamical systems with *quasi-periodic limit cycles* show a time evolution that can be decomposed into different periodic parts without a unique periodic regime. The corresponding time series consist of periodic parts of oscillation without a common structure. Nevertheless, closely starting trajectories remain close to each other during time evolution. The third class contains dynamical systems with *chaotic attractors* that are *nonperiodic*, with an *exponential dependence on initial conditions* for *bounded orbits*. A famous example is the chaotic attractor of a Lorenz system (Lorenz 1963) simulating the chaotic development of weather caused by local events, which cannot be forecast in the long run ("butterfly effect") (figure 3.4b).

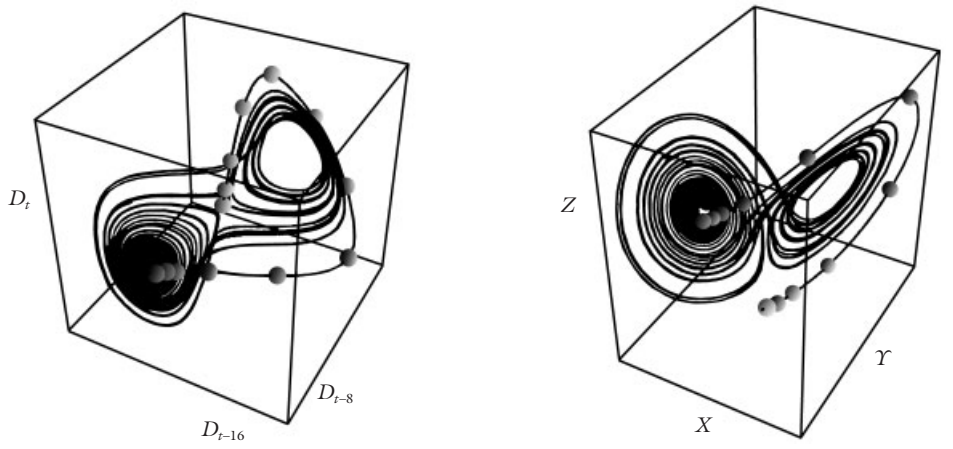

Figure3.4a: The reconstructed trajectory of a measured series (fig. 3.3) in an embedding phase space of three dimensions with time lag
Figure 3.4b: The trajectory of the original phase space of the chaotic system (Kaplan 1995: 310)

3 Dynamical Systems and Time-series Analysis

We have started by seeing the kind of mathematical equations of dynamical systems required to derive their patterns of behavior; the latter have been characterized by time series and attractors in phase spaces, such as fixed points, limit cycles, and chaos. This *top-down approach* is typically theoretical: we use our understanding of real systems to write dynamical equations. In empirical practice, however, we must take the opposite *bottom-up approach* and start with finite sequences of measurements, i.e. finite time series, in order to find appropriate equations of mathematical models with predictions that can be compared with measurements made in the field of application.

Measurements are often contaminated by unwanted noise, which must be separated from the signals of specific interest. Moreover, in order to forecast the behavior of a system, the development of its future states must be reconstructed in a corresponding phase space from a finite sequence of measurements. So *time-series analysis* is an immense challenge in different fields of research such as climatic data in meteorology, ECG-signals in cardiology, EEG-data in brain research, or economic data of business cycles in economics.

The goal for this kind of *time-series analysis* is comparable to constructing *a computer program* without any knowledge of the real system from which the data come. As a black box, the computer program would take the measured data as input and provide as output a mathematical model describing the data. But, in this case, it is difficult to identify the meaning of components in the mathematical model without understanding the dynamics of the real systems. Thus, the top-down and bottom-up approach, model-building and time-series analysis, expert knowledge in the fields of application, and mathematical and programming skills, must all be integrated in an *interdisciplinary research strategy*.

In practice, only a time series of a single (one-dimensional) measured variable is often given, although the real system is multidimensional. The aim of forecasting is to predict the future evolution of this variable. According to Takens' theorem (1981), in nonlinear, deterministic, and chaotic systems, it is possible to determine the structure of the *multidimensional dynamic system* from the measurement of a *single dynamical variable* (figure 3.3).

Takens' method results in the construction of a multidimensional *embedding phase space for measured data* (figure 3.4a) with a certain time lag in which the dynamics of attractors is similar

to the orbits in the phase space of the chaotic system (figure 3.4b).

The disadvantage of Takens' theorem is that it does not detect and prove the existence of a chaotic attractor. It only provides features of an attractor from measured data, if the existence of the attractor is already guaranteed (Grassberger & Procaccia 1983). The *dimension of an attractor* can be determined by a *correlation integral* defining the different frequency with which a region in an attractor is visited by the orbits. Thus, the correlation integral also provides a method to study the *degrees of periodicity and aperiodicity* of orbits and measured time series.

The *Lyapunov spectrum* shows us the *dependence of dynamics from initial data*. The so-called Lyapunov exponents measure the averaged exponential rates of divergence or convergence of neighboring orbits in phase space. If the largest Lyapunov exponent is positive, the attractor is *chaotic*, and the initial small difference between two trajectories will diverge exponentially (figure 3.2). If the largest exponent is zero and the rest is negative, the attractor is a *periodic limit cycle*. If there is more than one exponent equal to zero, the rest being negative, the behavior is *quasi-periodic*. If the exponents are all negative, the attractor is a *fixed point*. In general, for *dissipative systems*, the sum of Lyapunov exponents is negative, despite the fact that some exponents could be positive.

4 Dynamical Systems in Nature and Society

Structures in nature and society can be explained by the *dynamics* and *attractors* of complex systems. They result from collective patterns of interacting elements that cannot be reduced to the features of single elements in a complex system. *Nonlinear interactions* in multicomponent ("complex") systems often have synergetic effects, which can neither be traced back to single causes nor be forecasted in the long run. The mathematical formalism of complex dynamical systems is taken from statistical physics. In general, the theory of complex dynamical systems deals with profound and striking analogies that have been discovered in the self-organized behavior of quite different systems in physics, chemistry, biology, and sociology. These multicomponent systems consist of many units like elementary particles, atoms, cells, or organisms. Properties of these elementary units, such as their position and momentum vectors, and their local interactions constitute the microscopic level of description (imagine the interacting molecules of a liquid or gas). The global state of the complex systems results from the collective configurations of the local multicomponent states. At the macroscopic level, there are few collective ("global") quantities like, for instance, pressure, density, temperature, and entropy characterizing observable collective patterns or figures of the units.

If the external conditions of a system are changed by varying certain control parameters (e.g., temperature), the system may undergo a change in its macroscopic global states at some threshold value. For instance, water as a complex system of molecules changes spontaneously from a liquid to a frozen state at the critical value of temperature with zero celsius. In physics, those transformations of collective states are called *phase transitions*. Obviously, they describe a change of self-organized behavior between the interacting elements of a complex system.

According to Landau and Lifshitz (1959), the suitable macrovariables characterizing this change of global order are denoted as "*order parameters*." In statistical mechanics the order transition of complex systems like fluids, gases, etc. is modeled by differential equations of the global state. A paradigmatic example is a ferromagnet consisting of many elementary atomic magnets ("dipoles"). The two possible local states of a dipole are represented by upwards- and downwards-pointing arrows. If the temperature ("*control parameter*") is annealed to the thermal equilibrium, in this case the Curie point, then the average distribution of upwards and downwards pointing dipoles ("*order parameter*") is spontaneously aligned in one regular direction (figure 3.5). This regular pattern corresponds to the macroscopic state of magnetization. Obviously, the emergence of magnetization is a self-organized behavior of atoms that is modeled by a phase transition of a certain order parameter,

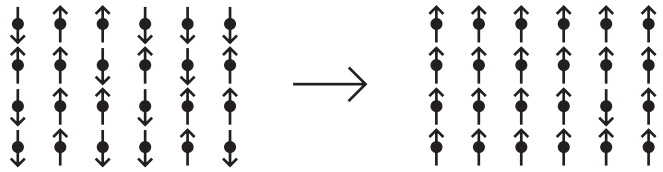

Figure 3.5: Phase transition in a 2-dimensional Ising model of a ferromagnet (Mainzer 1997: 134)

the average distribution of upwards and downwards pointing dipoles.

Landau's scheme of phase transitions cannot be generalized to all cases of phase transitions. A main reason for its failure results from an inadequate treatment of fluctuations, which are typical for many multicomponent systems. Nevertheless, Landau's scheme can be used as a heuristic device to deal with several *non-equilibrium transitions*. In this case, a complex system is driven away from equilibrium by increasing energy (not decreasing energy, as in the case of equilibrium transitions like freezing water or magnetizing ferromagnets). The phase transitions of nonlinear dissipative complex systems far from thermal equilibrium can be modeled by several mathematical methods (Haken 1983, Mainzer 1997, Glansdorff & Prigogine 1971).

As an example, consider a *solid-state laser*. This consists of a set of laser-active atoms embedded in a solid-state configuration. The laser end-faces act as mirrors and serve two purposes: they select light modes in axial direction and with discrete frequencies. If the laser atoms are excited only weakly by external sources ("*control parameters*"), the laser acts as an ordinary lamp. The atoms, independently of each other, emit wavetracks with random phases. The atoms, visualized as oscillating dipoles, are oscillating completely at random. If the level of excitement is further increased, the atomic dipoles spontaneously oscillate in phase, although they are excited completely at random. Obviously, the atoms show a *self-organized behavior of great regularity*. The extraordinary coherence of laser light results from the collective cooperation of the atomic dipoles.

The laser shows *features of phase transitions*. Order parameters describe mode amplitudes of the light field becoming *unstable* at a critical value of pumping. These slowly varying amplitudes now "*slave*," as Haken (1983) claimed, the atomic system during a critical transition. The atoms have to "obey" the orders of order parameters. This mathematical scheme has a very comfortable consequence: it is not necessary (and not possible) to compute all microstates of atoms in a complex system; just find the few macroscopic order parameters, and you understand the dynamics of a complex system.

Actually, the corresponding equations describe a competition of several order parameters among each other. The atoms will then obey that order parameter that wins the competition. A typical example is a *Bénard experiment* analyzing the emergence of convection rolls in a fluid layer at a critical value of a control parameter (temperature). The layers of the atmosphere provide further examples. In this case, the order parameters correspond to the two possible rolling directions: "left" or "right" of the convection rolls. During the *phase transition* of increasing temperature it cannot be forecast which of the two possible order parameters will win the competition, because it depends on tiny initial fluctuations on the molecular level. Thus, this phase transition corresponds to a *spontaneous symmetry breaking* of two possible orders. *Fluctuations* are the driving forces of the system's evolution.

Simplifying, we may say that old structures become unstable, broken down by changing control parameters, and *new structures* and attractors are achieved. If, for example, the fluid of a stream is driven further and further away from thermal equilibrium, for example by increasing fluid velocity (control parameter), then fluid patterns of increasing complexity emerge from *vortices of fixed points*, *periodic* and *quasi-periodic oscillations* to *chaotic turbulence*.

More mathematically, stochastic nonlinear differential equations (e.g. Fokker–Planck equations, Master equations) are employed to model the dynamics of complex systems. The dominating

order parameters are founded by the adiabatic elimination of fast-relaxing variables of these equations. The reason is that the relaxation time of unstable modes (order parameters) is very long, compared to the fast-relaxing variables of stable ones, which can therefore be neglected. Thus, this concept of self-organization can be illustrated by a quasi-biological slogan: long-living systems dominate short-living systems.

Dynamical systems and their *phase transitions* deliver a successful formalism to model the *emergence of order in nature and society*. But these methods are not reduced to special laws of physics, although their mathematical principles were first discovered and successfully applied in physics. Methodologically, there is *no physicalism*, but an interdisciplinary approach to explain the increasing complexity and differentiation of forms by phase transitions. The question is how to select, interpret, and quantify the appropriate variables of dynamical models. Let us consider a few examples.

Thermodynamic self-organization is not sufficient to explain the emergence of life (see also Chapter 14 in this volume, CYBERNETICS). As nonlinear mechanism of genetics we use the autocatalytic process of genetic self-replication. The evolution of new species by mutation and selection can be modeled by nonlinear stochastic equations of second-order non-equilibrium phase transitions. Mutations are mathematized as "fluctuating forces" and selections as "driving forces." Fitness degrees are the order parameters dominating the phase transitions to new species. During evolution a sensible network of equilibria between populations of animals and plants has developed. The nonlinear Lotka–Volterra equations (Lotka 1925, Volterra 1931) model the ecological equilibrium between prey and predator populations which can be represented by oscillating time series of population growth or limit cycles around points of stability. Open dissipative systems of ecology may become unstable because of local perturbations, e.g., pollution of the atmosphere, leading to global chaos of the atmosphere in the sense of the *butterfly effect*.

In cardiology, the heart is modeled as a complex dynamical system of electrically interacting cells producing collective patterns of beating, which are then represented by time series of ECG signals or orbits in a phase space. There is no commanding cell, but an attractor of collective behavior ("order parameter") dominating the beating regime of the heart from healthy oscillations to dangerous chaos.

In brain research, the brain is considered as a complex dynamical system of firing and nonfiring neurons, self-organizing in macroscopic patterns of cell assemblies through neurochemical interactions. Their dynamical attractors are correlated with states of perception, motion, emotion, thoughts, or even consciousness. There is no "mother neuron" that can feel, think, or at least coordinate the appropriate neurons. The famous *binding problem* of pixels and features in a perception is explained by clusters of synchronously firing neurons dominated by learned attractors of brain dynamics.

The self-organization of complex systems can also be observed in social groups. If a group of workers is directed by another worker, the so-called foreman, then we get an organized behavior to produce some product that is by no means self-organized. Self-organization means that there are no external orders from a foreman, but that the workers work together by some kind of mutual understanding, each one doing his job according to a collective concept dominating their behavior.

In a political community, collective trends or majorities of opinions can be considered as order parameters produced by mutual discussion and interaction of the people in a more or less "heated" situation. They can even be initiated by some few people in a critical and unstable ("revolutionary") situation of the whole community. There may be a competition of order concepts during heavy fluctuations. The essential point is that the winning concept of order will dominate the collective behavior of the people. Thus, there is a kind of feedback: the collective order of a complex system is generated by the interactions of its elements ("*self organization*"). On the other hand, the behavior of the elements is dominated by the collective order. People have their individual will to influence collective trends of society. But, they are also driven by attractors of collective behavior.

In classical economics, an economy was believed to be a conservative equilibrium system.

According to Adam Smith (1976), the market is self-organized by an "invisible hand," tending to the equilibrium of supply and demand. In the age of globalization, markets are open, non-equilibrium systems at the edge of chaos (in the technical sense of the word seen above), with sensible dependence on local perturbations (*butterfly effect*). The time series of stock markets and business cycles are examples of economic signals.

Another application of social dynamics is the behavior of car drivers. In automobile traffic systems, a phase transition from nonjamming to jamming phases depends on the averaged car density as control parameter. At a critical value, fluctuations with fractal or self-similar features can be observed. The term self-similarity states that the time series of measured traffic flow looks the same on different time scales, at least from a qualitative point of view with small statistical deviations. In the theory of complex systems, self-similarity is a (not sufficient) hint at chaotic dynamics. These signals can be used by traffic guiding systems.

5 Dynamical, Information, and Computational Systems

Dynamical systems can be characterized by *information* and *computational concepts*. A dynamical system can be considered as an *information processing machine*, computing a present state as output from an initial state of input. Thus, the *computational efforts* to determine the states of a system characterize the *complexity of a dynamical system*. The *transition from regular to chaotic system*s corresponds to increasing computational problems, according to increasing degrees in the *computational theory of complexity* (see Chapter 1 in this volume, COMPUTATION). In statistical mechanics, the *information flow* of a dynamical system describes the intrinsic evolution of statistical correlations. In *chaotic systems* with sensitivity to the initial states, there is an increasing *loss of information* about the initial data, according to the decay of correlations between the entire past and future states of the system. In general, *dynamical systems* can be considered as *deterministic*, *stochastic*, or *quantum computers*, computing information about present or future states from initial conditions by the corresponding dynamical equations. In the case of quantum systems, the binary concept of information is replaced by *quantum information* with superposition of binary digits. Thus, *quantum information* only provides probabilistic forecasts of future states.

The complex system approach offers a research program to bridge the gap between *brain research* and *cognitive science*. In a famous metaphor, Leibniz compared the machinery of a human brain and body with the machinery of a mill that can be explored inside and observed in its behavior. In modern brain research, the interacting cogs of the mill are the firing and nonfiring neurons which could be technically constructed by a neural net. If the human brain is considered as a complex dynamical system, then emergence of mental states can be modeled by phase transitions of macroscopic order parameters which are achieved by collective nonlinear interactions of neurons, but which are not reducible to microscopic states of the system: A single neuron cannot think or feel. The complex system approach is an empirical research program that can be specified and tested in appropriate experimental applications to understand the dynamics of the human cognitive system. Further on, it gives heuristic devices to construct artificial systems with cognitive features in robotics (see Chapters 8, 13–16 in this volume).

In a dramatic step, the complex systems approach has been enlarged from neural networks to *global computer networks* like the *World Wide Web*. The internet can be considered as a *complex open computer network* of autonomous nodes (hosts, routers, gateways, etc.), self-organizing without central control mechanisms. The information traffic is constructed by information packets with source and destination addresses. *Routers* are nodes of the network determining the local path of each packet by using local routing tables with cost metrics for neighboring routers. A router forwards each packet to a neighboring router with lowest costs to the destination. As a router can only deal with one packet, other arriving packets at a certain time must be stored in a buffer. If more packets arrive than a buffer can

store, the router discards the overflowed packets. Senders of packets wait for confirmation message from the destination host. These buffering and resending activities of routers can cause *congestion in the internet*. A *control parameter* of data density is defined by the propagation of congestion from a router to neighboring routers and dissolution of the congestion at each router. The cumulative distribution of congestion duration is an *order parameter of phase transition*. At a critical point, when the congestion propagation rate is equal to congestion dissolution, *fractal and chaotic features* can be observed in data traffic. Congested buffers behave in surprising analogy to infected people. If a buffer is overloaded, it tries to send packets to the neighboring routers. Therefore the congestion spreads spatially. On the other hand, routers can recover when the congestion from and to the own subnet are lower than the service rate of the router. That is not only an illustrative metaphor, but hints at *nonlinear mathematical models* describing true *epidemic processes* like malaria extension as well as the dynamics of routers. Computer networks are *computational ecologies*. The capability to manage the complexity of modern societies depends decisively on effective communication networks.

The transformation of the internet into a system with *self-organizing features of learning and adapting* is not merely a metaphor. *Information retrieval* is already realized by *neural networks* adapting to the information preferences of a human user with *synaptic plasticity*. In sociobiology, we can learn from populations of ants and termites how to organize traffic and information processing by *swarm intelligence*. From a technical point of view, we need intelligent programs distributed in the nets. There are already more or less intelligent virtual organisms ("*agents*"), learning, self-organizing, and adapting to our individual preferences of information, selecting our e-mails, preparing economic transactions, or defending against attacks by hostile computer viruses, like the immune system of our body. Complexity of global networking not only means increasing numbers of PCs, workstations, servers, and supercomputers interacting via data traffic in the internet. Below the complexity of a PC, low-power, cheap, and smart devices are distributed in the intelligent environments of our everyday world. Like GPS (the Global Position System) in car traffic, things in everyday life could interact *telematically* by sensors. The real power of the concept does not come from any one of these single devices. In the sense of *complex systems*, the power emerges from the collective interaction of all of them. For instance, the optimal use of energy could be considered as a macroscopic *order parameter* of a household realized by the *self-organizing use* of different household goods according to less consumption of electricity during special time-periods with cheap prices. The processors, chips, and displays of these smart devices don't need a user interface like a mouse, windows, or keyboards, but just a pleasant and effective place to get things done. *Wireless computing devices* on small scales become more and more invisible to the user. Ubiquitous computing enables people to live, work, use, and enjoy things directly without being aware of their computing devices.

What are the *human perspectives* in these developments of *dynamical*, *information*, and *computational systems*? Modern societies, economies, and information networks are highly dimensional systems with a complex nonlinear dynamics. From a methodological point of view, it is a challenge to improve and enlarge the instruments of modelization (cf. sections 1–3 above) from low- to high-dimensional systems. Modern systems science offers an *interdisciplinary methodology* to understand the typical features of self-organizing dynamics in nature and society. As nonlinear models are applied in different fields of research, we gain general insights into the predictable horizons of oscillatory chemical reactions, fluctuations of species, populations, fluid turbulence, and economic processes. The emergence of sunspots, for instance, which was formerly analyzed by statistical time-series methods, is by no means a random activity. It can be modeled by a nonlinear chaotic system with several characteristic periods and a strange attractor, allowing bounded forecasts of the variations. In nonlinear models of public opinion formation, for instance, we may distinguish a predictable stable state before public voting (*bifurcation*) when neither of two possible opinions is preferred, a short interval of bifurcation when tiny unpredictable fluctuations

may induce abrupt changes, and a transition to a stable majority. The situation can be compared to growing air bubbles in turbulently boiling water: When a bubble has become big enough, its steady growth on its way upward is predictable. But its origin and early growth is a question of random fluctuation. Obviously, nonlinear modeling explains the difficulties of the modern sibyls of demoscopy.

Today, nonlinear forecasting models don't always deliver better and more efficient predictions than the standard linear procedures. Their main advantage is the explanation of the actual nonlinear dynamics in real processes, the identification and improvement of local horizons with short-term predictions. But first of all the phase space and an appropriate dynamical equation governing a time series of observations must be reconstructed to predict future behavior by solving that equation. Even in the natural sciences, it is still unclear whether appropriate equations for complex fields such as earthquakes can be derived. We may hope to set up a list in a computer memory with typical nonlinear equations whose coefficients can be automatically adjusted for the observed process. Instead, to make an exhaustive search for all possible relevant parameters, a learning strategy may start with a crude model operating over relatively short times, and then specify a smaller number of parameters in a relatively narrow range of values. An improvement of short-term forecasting has been realized by the learning strategies of neural networks. On the basis of learned data, neural nets can weight the input data and minimize the forecasting errors of short-term stock quotations by self-organizing procedures. So long as only some stock-market advisers use this technical support, they may do well. But if all agents in a market use the same learning strategy, the forecasting will become a self-defeating prophecy. The reason is that human societies are not complex systems of molecules or ants, but the result of highly *intentional acting beings* with a greater or lesser amount of free will. A particular kind of *self-fulfilling prophecy* is the Oedipus effect: like the legendary Greek king, people try, in vain, to change their future as forecasted to them.

From a macroscopic viewpoint we may, of course, observe single individuals contributing with their activities to the collective macrostate of society representing cultural, political, and economic order (*order parameters*). Yet, macrostates of a society, of course, don't simply average over its parts. Its order parameters strongly influence the individuals of the society by orientating (*enslaving*) their activities and by activating or deactivating their attitudes and capabilities. This kind of feedback is typical for complex dynamical systems. lf the control parameters of the environmental conditions attain certain critical values due to internal or external interactions, the macrovariables may move into an unstable domain out of which highly divergent alternative paths are possible. Tiny unpredictable microfluctuations (e.g., actions of a few influential people, scientific discoveries, new technologies) may decide which of the diverging paths in an unstable state of bifurcation society will follow. So, the paradigm of a centralized control must be given up by the insights in the self-organizing dynamics of highly dimensional systems. By detecting global trends and the order parameters of complex dynamics, we have the chance of implementing favorite tendencies. By understanding complex systems we can make much more progress in evaluating our information technologies and choosing our next steps. Understanding complex systems supports deciding and acting in a complex world.

References

Arnold, V. I. 1963. "Small denominators II. Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally-periodic motions under a small perturbation of the Hamiltonian." *Russian Mathematical Surveys* 18(5). [Proof of KAM theorem; graduate level.]

Glansdorff, P. and Prigogine, I. 1971. *Thermodynamic Theory of Structures, Stability and Fluctuations.* New York: Wiley. [Basic textbook of dissipative structures; graduate level.]

Grassberger, P. and Procaccia, I. 1983. "Characterization of strange attractors." *Physical Revue Letters* 50: 346–9. [Theorem of characterizing chaotic attractor in time series; graduate level.]

Haken, H. 1983. *Synergetics: Nonequilibrium Phase Transitions and Self-organization in Physics,*

Chemistry, and Biology, 3rd ed. Berlin: Springer. [Basic textbook on synergetics; undergraduate level.]

Holland, J. H. 1992. *Adaption in Natural and Artifical Systems.* Cambridge MA: MIT Press. [An introduction; undergraduate level.]

Kaplan, D. and Glass, L. 1995. *Understanding Nonlinear Dynamics.* New York: Springer. [Introduction to nonlinear dynamics; undergraduate level.]

Kolmogorov, A. N. 1954. "On conservation of conditionally-periodic motions for a small change in Hamilton's function." *Dokl. Akad. Nauk SSSR* 98: 527–30. [Proof of KAM theorem: graduate level.]

Landau, L. D. and Lifshitz, E. M. 1959. *Course of Theoretical Physics*, vol. 6: *Fluid Mechanics.* London: Pergamon Press. [Famous fluid mechanics textbook; graduate level.]

Laplace, P.-S. de. 1814. *Essai philosophique sur les probabilités.* Paris. [Historically important essay; undergraduate level.]

Lorenz, E. N. 1963. "Deterministic nonperiodic flow." *Journal of Atmospheric Science* 20: 130–41. [Detection of Lorenz's attractor: graduate level.]

Lotka, A. J. 1925. *Elements of Mathematical Biology.* New York: Dover. [Historically important ecological systems science textbook; undergraduate level.]

Mainzer, K. 1997. *Thinking in Complexity: The Complex Dynamics of Matter, Mind, and Mankind.* 3rd ed. Berlin: Springer. [Interdisciplinary and philosophical introduction to complex systems; undergraduate level.]

Moser, J. 1967. "Convergent series expansions of quasi-periodic motions." *Mathematical Annals* 169: 163. [Proof of KAM theorem; graduate level.]

Poincaré, H. 1892–3. *Les Méthodes nouvelles de la mécanique célèste I–III.* Paris: Gauthier-Villars. [Historically important source of chaos theory; graduate level.]

Smith, A. 1976. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Chicago: University of Chicago Press. [Historically important source on economic self-organization; undergraduate level.]

Takens, F. 1981. "Detecting strange attractors in turbulence." In D. A. Rand and L. S. Young, eds., *Dynamical Systems and Turbulence.* Berlin: Springer, pp. 336–86. [Takens' theorem for detecting chaotic attractors in time series; graduate level.]

Volterra, V. 1931. *Leçons sur la théorie mathématique de la lutte pour la vie.* Paris. [Historically important ecological systems science textbook; graduate level.]

Chapter 4

Information

Luciano Floridi

1 Introduction

Information "can be said in many ways," just as being can (Aristotle, *Metaphysics* Γ.2), and the correlation is probably not accidental. Information, with its cognate concepts like computation, data, communication, etc., plays a key role in the ways we have come to understand, model, and transform reality. Quite naturally, information has adapted to some of being's contours.

Because information is a multifaceted and polyvalent concept, the question "what is information?" is misleadingly simple, exactly like "what is being?" As an instance of the Socratic question "*ti esti* . . . ?," it poses a fundamental and complex problem, intrinsically fascinating and no less challenging than "what is truth?," "what is virtue?," "what is knowledge?," or "what is meaning?" It is not a request for dictionary explorations but an ideal point of intersection of philosophical investigations, whose answers can diverge both because of the conclusions reached and because of the approaches adopted. Approaches to a Socratic question can usually be divided into three broad groups: reductionist, antireductionist, and nonreductionist. Philosophical theories of information are no exception.

Reductionists support the feasibility of a "unified theory of information" (UTI, see the UTI website for references), general enough to capture all major concepts of information (from Shannon's to Baudrillard's, from genetic to neural), but also sufficiently specific to discriminate between semantic nuances. They attempt to show that all kinds of information are ultimately reducible conceptually, genetically, or genealogically to some *Ur*-concept, the mother of all instances. The development of a systematic UTI is a matter of time, patience, and ingenuity. The ultimate UTI will be hierarchical, linear (even if probably branching), inclusive, and incompatible with any alternative model.

Reductionist strategies are unlikely to succeed. Several surveys have shown no consensus or even convergence on a single, unified definition of information (see for example Braman 1989, Losee 1997, Machlup 1983, NATO 1974, 1975, 1983, Schrader 1984, Wellisch 1972, Wersig & Neveling 1975). This is hardly surprising. Information is such a powerful and flexible concept and such a complex phenomenon that, as an *explicandum*, it can be associated with several explanations, depending on the level of abstraction adopted and the cluster of requirements and desiderata orientating a theory. Claude Shannon (1993a: 180), for one, was very cautious:

> The word "information" has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently

> useful in certain applications to deserve further study and permanent recognition. *It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.* [italics added]

At the opposite end, antireductionists stress the multifarious nature of the concept of information and of the corresponding phenomena. They defend the radical irreducibility of the different species to a single stem, objecting especially to reductionist attempts to identify Shannon's quantitative concept of information as the required *Ur*-concept and to ground a UTI on the mathematical theory of communication. Antireductionist strategies are essentially negative and can soon become an impasse rather than a solution. They allow specialized analyses of the various concepts of information to develop independently, thus avoiding the vague generalizations and mistaken confusions that may burden UTI strategies. But their fragmented nominalism remains unsatisfactory insofar as it fails to account for the ostensible connections permeating and influencing the various ways in which information *qua* information "can be said." *Connections*, mind, not Wittgensteinian *family resemblances*. The genealogical analogy would only muddy the waters here, giving the superficial impression of having finally solved the difficulty by merely hiding the actual divergences. The die-hard reductionist would still argue that all information concepts descend from the same *family*, while the unrepentant antireductionist would still object that we are facing mere *resemblances*, and that the various information concepts truly have different roots.

Nonreductionists seek to escape the dichotomy between reductionism and antireductionism by replacing the reductionist hierarchical model with a distributed network of connected concepts, linked by mutual and dynamic influences that are not necessarily genetic or genealogical. This "hypertextual analysis" can be centralized in various ways or completely decentralized and perhaps multicentered.

According to decentralized or multicentered approaches, there is no key concept of information. More than one concept is equally important, and the "periphery" plays a counterbalancing role. Depending on the orientation, information is seen as *interpretation*, *power*, *narrative*, *message* or *medium*, *conversation*, *construction*, a *commodity*, and so on. Thus, philosophers like Baudrillard, Foucault, Lyotard, McLuhan, Rorty, and Derrida are united by what they dismiss, if not challenge: the predominance of the factual. For them information is not in, from, or about reality. They downplay the *aboutness* of information and bend its referential thrust into a self-referential circle of hermeneutical communication. Their classic target is Cartesian foundationalism seen as the clearest expression of a hierarchical and authoritarian approach to the genesis, justification, and flow of information. Disoriented, they mistake it (Cartesian foundationalism) as the only alternative to their fully decentralized view.

Centralized approaches interpret the various meanings, uses, applications, and types of information as a system gravitating around a core notion with theoretical priority. The core notion works as a hermeneutical device that influences, interrelates, and helps to access other notions. In metaphysics, Aristotle held a similar view about being, and argued in favor of the primacy of the concept of *substance*. In the philosophy of information, this "substantial" role has long been claimed by *factual* or *epistemically oriented* semantic information. The basic idea is simple. In order to understand what information is, the best thing to do is to start by analyzing it in terms of the knowledge it can yield about its reference (the "abouted"). This epistemic approach is not without competitors. Weaver (1949), for example, supported a tripartite analysis of information in terms of (1) technical problems concerning the quantification of information and dealt with by Shannon's theory; (2) semantic problems relating to meaning and truth; and (3) what he called "influential" problems concerning the impact and effectiveness of information on human behavior, which he thought had to play an equally important role. Moreover, in pragmatic contexts, it is common to privilege a view of information as primarily a resource for decision-making processes. One of the tasks of this chapter is to show how in each case the centrality of epistemically oriented semantic information is presupposed rather than replaced.

We are now well placed to look at the structure of this chapter. In the following pages the question "what is information?" is approached from a nonreductionist and epistemically centralized perspective. In section 2, the concept of semantic information is reviewed assuming that factual information is the most important and influential sense in which information *qua* information "can be said." However, no attempt is made to reduce all other concepts to factual information. Factual information is like the capital of the informational archipelagos, crucially positioned to provide both a clear grasp of what information is and a privileged gateway to other important concepts that are interconnected but not necessarily reducible to a single *Ur*-concept. To show this in practice and to enrich our understanding of what else information may be, we shall look at two neighboring areas of great importance. Section 3 summarizes the mathematical theory of communication, which studies the statistical behavior of uninterpreted data, a much-impoverished concept of information. Section 4 outlines some important philosophical programs of research that investigate a more enriched concept of semantic information. Space constraints prevent discussion of several other important concepts of information, but some of them are briefly mentioned in the conclusion.

2 Semantic Information

In this section, a general definition of semantic information is introduced, followed by a special definition of factually oriented semantic information. The contents of the section are based on Floridi forthcoming *a* and *c*. The approach is loosely connected with the methodology developed in situation logic (see section 3.2).

2.1 Semantic information as content

Information is often used in connection with communication phenomena to refer to objective (in the sense of mind-independent or external, and informee-independent) *semantic contents*. These can be of various size and value, formulated in a range of codes and formats, embedded in physical implementations of different kinds. They can variously be produced, processed, communicated, and accessed. The *Cambridge Dictionary of Philosophy*, for example, defines information thus:

> an objective (mind independent) entity. It can be generated or carried by messages (words, sentences) or by other products of cognizers (interpreters). Information can be encoded and transmitted, but the information would exist independently of its encoding or transmission.

Examples of information in this broad sense are this *Guide*, Edgar Allan Poe's *The Raven*, Verlaine's *Song of Autumn*, the Rosetta Stone and the movie *Fahrenheit 451*.

Over the last three decades, many analyses have converged on a General Definition of Information (GDI) as semantic content in terms of *data + meaning* (see Floridi forthcoming *a* for extended bibliography):

GDI) σ is an instance of information, understood as objective semantic content, if and only if:
GDI.1) σ consists of n *data* (d), for $n \geq 1$;
GDI.2) the data are *well-formed* (wfd);
GDI.3) the wfd are *meaningful* (mwfd = δ).

GDI has become an operational standard especially in fields that treat data and information as reified entities (consider, for example, the now common expressions "data mining" and "information management"). Examples include Information Science; Information Systems Theory, Methodology, Analysis, and Design; Information (Systems) Management; Database Design; and Decision Theory. Recently, GDI has begun to influence the philosophy of computing and information (Floridi 1999 and Mingers 1997).

According to GDI, information can consist of different types of data δ. Data can be of four types (Floridi 1999):

δ.1) *primary data*. These are the principal data stored in a database, e.g. a simple array of numbers. They are the data an information-management system is

generally designed to convey to the user in the first place.

δ.2) *metadata*. These are secondary indications about the nature of the primary data. They describe properties such as location, format, updating, availability, copyright restrictions, and so forth.

δ.3) *operational data*. These are data regarding usage of the data themselves, the operations of the whole data system and the system's performance.

δ.4) *derivative data*. These are data that can be extracted from δ.1–δ.3, whenever the latter are used as sources in search of patterns, clues, or inferential evidence, e.g. for comparative and quantitative analyses (*ideometry*).

GDI indicates that information cannot be dataless, but it does not specify which types of data constitute information. This *typological neutrality* (TyN) is justified by the fact that, when the apparent absence of data is not reducible to the occurrence of *negative* primary data, what becomes available and qualifies as information is some further nonprimary information μ about σ constituted by some nonprimary data δ.2–δ.4. For example, if a database query provides an answer, it will provide at least a *negative* answer, e.g. "no documents found." If the database provides no answer, either it fails to provide any data at all, in which case no specific information σ is available, or it can provide some data δ to establish, for example, that it is running in a loop. Likewise, silence, as a reply to a question, could represent negative information, e.g. as implicit assent or denial, or it could carry some nonprimary information μ, e.g. the person has not heard the question.

Information cannot be dataless but, in the simplest case, it can consist of a single datum (*d*). A datum is reducible to just a lack of uniformity between two signs. So our definition of a datum (*Dd*) is:

Dd) $d = (x \neq y)$, where the *x* and the *y* are two uninterpreted variables.

The dependence of information on the occurrence of syntactically well-formed data, and of data on the occurrence of differences variously implementable physically, explain why information can be decoupled from its support. Interpretations of this support-independence vary radically because Dd leaves underdetermined not only the logical type to which the relata belong (see TyN), but also *the classification* of the relata (*taxonomic neutrality*), *the kind of support* required for the implementation of their inequality (*ontological neutrality*), and the dependence of their semantics on a producer (*genetic neutrality*).

Consider the *taxonomic neutrality* (TaN) first. A datum is usually classified as the entity exhibiting the anomaly, often because the latter is perceptually more conspicuous or less redundant than the background conditions. However, the relation of inequality is binary and symmetric. A white sheet of paper is not just the necessary background condition for the occurrence of a black dot as a datum, it is a constitutive part of the datum itself, together with the fundamental relation of inequality that couples it with the dot. Nothing is a datum *per se*. Being a datum is an external property. GDI endorses the following thesis:

TaN) A datum is a relational entity.

So, no data without relata, but GDI is neutral with respect to the identification of data with specific relata. In our example, GDI refrains from identifying either the black dot or the white sheet of paper as the datum.

Understood as relational entities, data are *constraining affordances*, exploitable by a system as input of adequate queries that correctly semanticize them to produce information as output. In short, information as content can also be described erotetically as *data + queries* (Floridi 1999). I shall return to this definition in section 3.2.

Consider now the *ontological neutrality* (ON). By rejecting the possibility of dataless information, GDI endorses the following modest thesis:

ON) No information without data representation.

Following Landauer and Bennett 1985 and Landauer 1987, 1991, and 1996, ON is often

interpreted materialistically, as advocating the impossibility of physically disembodied information, through the equation "representation = physical implementation":

ON.1) No information without physical implementation.

ON.1 is an inevitable assumption when working on the physics of computation, since computer science must necessarily take into account the physical properties and limits of the data carriers. Thus, the debate on ON.1 has flourished especially in the context of the philosophy of quantum computing (see Landauer 1991, Deutsch 1985, 1997; Di Vincenzo & Loss 1998; Steane 1998 provides a review). ON.1 is also the ontological assumption behind the Physical Symbol System Hypothesis in AI and Cognitive Science (Newell & Simon 1976). But ON, and hence GDI, does not specify whether, ultimately, the occurrence of every discrete state necessarily requires a *material* implementation of the data representations. Arguably, environments in which all entities, properties, and processes are ultimately noetic (e.g. Berkeley, Spinoza), or in which the material or extended universe has a noetic or non-extended matrix as its ontological foundation (e.g. Pythagoras, Plato, Descartes, Leibniz, Fichte, Hegel), seem perfectly capable of upholding ON without necessarily embracing ON.1. The relata in Dd could be monads, for example. Indeed, the classic realism debate can be reconstructed in terms of the possible interpretations of ON.

All this explains why GDI is also consistent with two other popular slogans this time favorable to the proto-physical nature of information and hence completely antithetic to ON.1:

ON.2) "*It from bit.* Otherwise put, every 'it' – every particle, every field of force, even the space-time continuum itself – derives its function, its meaning, its very existence entirely – even if in some contexts indirectly – from the apparatus-elicited answers to yes-or-no questions, binary choices, *bits.* 'It from bit' symbolizes the idea that every item of the physical world has at bottom – a very deep bottom, in most instances – an immaterial source and explanation; that which we call reality arises in the last analysis from the posing of yes–no questions and the registering of equipment-evoked responses; in short, that all things physical are information-theoretic in origin and that this is a *participatory universe*" (Wheeler 1990, 5);

and

ON.3) "[information is] a name for the content of what is exchanged with the outer world as we adjust to it, and make our adjustment felt upon it." (Wiener 1954, 17). "Information is information, not matter or energy. No materialism which does not admit this can survive at the present day" (Wiener 1961: 132).

ON.2 endorses an information-theoretic, metaphysical monism: the universe's essential nature is digital, being fundamentally composed of information as data instead of matter or energy, with material objects as a complex secondary manifestation (a similar position has been defended more recently in physics by Frieden 1998, whose work is based on a Platonist perspective). ON.2 may but does not have to endorse a strictly computational view of information processes. ON.3 advocates a more pluralistic approach along similar lines. Both are compatible with GDI.

A final comment concerning GDI.3 can be introduced by discussing a fourth slogan:

ON.4) "In fact, what we mean by information – the elementary unit of information – is a difference which makes a difference." (Bateson 1973: 428)

ON.4 is one of the earliest and most popular formulations of GDI (see for example Franklin 1995, 34 and Chalmers 1997: 281; note that the formulation in MacKay 1969, that is, "information is a *distinction* that makes a difference," predates Bateson's and, although less memorable, is more accurate). A "difference" is just a discrete state (that is, a datum), and "making a

difference" simply means that the datum is "meaningful," at least potentially.

Finally, let us consider the semantic nature of the data. How data can come to have an assigned meaning and function in a semiotic system in the first place is one of the hardest problems in semantics. Luckily, the point in question here is not *how* but *whether* data constituting information as semantic content can be meaningful *independently* of an informee. The *genetic neutrality* (GeN) supported by GDI states that:

GeN) δ can have a semantics *independently* of any informee.

Before the discovery of the Rosetta Stone, Egyptian hieroglyphics were already regarded as information, even if their semantics was beyond the comprehension of any interpreter. The discovery of an interface between Greek and Egyptian did not affect the semantics of the hieroglyphics but only its accessibility. This is the weak, conditional-counterfactual sense in which GDI. 3 speaks of meaningful data being embedded in information carriers informee-independently. GeN supports the possibility of *information without an informed subject*, to adapt a Popperian phrase. Meaning is not (at least not only) in the mind of the user. GeN is to be distinguished from the stronger, realist thesis, supported for example by Dretske (1981), according to which data could also have their own semantics independently of an intelligent *producer/informer*. This is also known as *environmental information*, and a typical example given is the series of concentric rings visible in the wood of a cut tree trunk, which may be used to estimate its age.

To summarize, GDI defines information broadly understood as semantic content comprised of syntactically well-formed and meaningful data. Its four types of neutrality (TyN, TaN, ON, and GeN) represent an obvious advantage, as they make GDI perfectly scalable to more complex cases and reasonably flexible in terms of applicability and compatibility. The next question is whether GDI is satisfactory when discussing the most important type of semantic information, namely factual information.

2.2 Semantic information as factual information

We have seen that semantic information is usually associated with communication. Within this context, the most important type of semantic information is *factual information*, which tells the informee something *about* something else, for example where a place is, what the time is, whether lunch is ready, or that penguins are birds. Factual information has a declarative (Kant's judicial) nature, is satisfactorily interpretable in terms of first-order, classic predicate logic, is correctly qualifiable alethically, and can be appropriately analyzed in the following form "*a*'s being (of type) *F* carries the information that *b* is *G*" (Dretske 1981, Barwise & Seligman 1997).

Does GDI provide a definition of factual information? Some philosophers (Barwise & Seligman 1997, Dretske 1981, Floridi forthcoming *a* and *c*. Grice 1989) have argued that it does not, because otherwise false information would have to count as a type of factual information, and there are no convincing reasons to believe it does, while there are compelling reasons to believe that it does not (for a detailed analysis see Floridi forthcoming *a*). As Dretske and Grice have put it: "*false* information and *mis*-information are not kinds of information – any more than decoy ducks and rubber ducks are kinds of ducks" (Dretske 1981: 45) and "False information is not an inferior kind of information; it just is not information" (Grice 1989: 371). Let us look at the problem in more detail.

The difficulty lies here with yet another important neutrality in GDI. GDI makes no comment on the truthfulness of data that may comprise information (*alethic neutrality* AN):

AN) Meaningful and well-formed data qualify as information, no matter whether they represent or convey a truth or a falsehood or have no alethic value at all.

Verlaine's *Song of Autumn* counts as information even if it does not make sense to ask whether it is true or false, and so does every sentence in *Old Moore's Almanac*, no matter how downright false. Information as purely semantic content is

completely decoupled from any alethic consideration (Colburn 2000, Fetzer forthcoming, and Fox 1983 can be read as defending this perspective). However, if GDI is also taken to define factual information, then

a) false information about the world (including contradictions), i.e. *misinformation*, becomes a genuine type of factual information;
b) tautologies qualify as factual information;
c) "it is true that *p*" where *p* can be replaced by any instance of genuine factual information, is no longer a redundant expression, e.g. "it is true" in the conjunction "'the earth is round' qualifies as information *and* it is true" cannot be eliminated without semantic loss; and finally
d) it becomes impossible to erase factual information semantically (we shall be more and more informed about *x*, no matter what the truth value of our data about *x* is).

None of these consequences is ultimately defensible, and their rejection forces a revision of GDI. "False" in "false information" is used attributively, not predicatively. As in the case of a false constable, false information is not factual information that is false, but not factual information at all. So "false information" is, like "false evidence," not an oxymoron, but a way of specifying that the informational contents in question do not conform to the situation they purport to map (or "to about"), and so fail to qualify as factual information. Well-formed and meaningful data may be of poor quality. Data that are incorrect (vitiated by errors or inconsistencies), imprecise (precision is a measure of the repeatability of the collected data), or inaccurate (accuracy refers to how close the average data value is to the actual value) are still data and may be recoverable. But, if they are not truthful, they can only amount to semantic content at best and misinformation at worst.

The special definition of information (SDI) needs to include a fourth condition about the positive alethic nature of the data in question:

SDI) σ is an instance of factual information if and only if:
SDI.1) σ consists of *n data* (*d*), for $n \geq 1$;
SDI.2) the data are *well-formed* (wfd);
SDI.3) the wfd are *meaningful* (mwfd = δ);
SDI.4) the δ are *truthful*.

Factual information encapsulates truthfulness, which does not contingently supervene on, but is necessarily embedded in it. And since information is "said primarily in factual ways," to put it in Aristotelian terms, false information can be dismissed as no factual information at all, although it can still count as information in the trivial sense of semantic content.

3 The Mathematical Theory of Communication

Some features of information are intuitively quantitative. Information can be *encoded*, *stored*, and *transmitted*. We also expect it to be *additive* and *non-negative*. Similar properties of information are investigated by the *mathematical theory of communication* (MTC) with the primary aim of devising efficient ways of encoding and transferring data.

MTC is not the only successful mathematical approach to information theory, but it certainly is the best and most widely known, and the one that has had the most profound impact on philosophical analyses. The name for this branch of probability theory comes from Shannon's seminal work (Shannon 1948, now Shannon & Weaver 1998). Shannon pioneered this field and obtained many of its principal results, but he acknowledged the importance of previous work done by other researchers at Bell laboratories, most notably Nyquist and Hartley (see Cherry 1978 and Mabon 1975). After Shannon, MTC became known as *information theory*, an appealing but unfortunate label, which continues to cause endless misunderstandings. Shannon came to regret its widespread popularity, and we shall avoid using it in this context.

This section outlines some of the key ideas behind MTC, with the aim of understanding the relation between MTC and the philosophy of information. The reader with no taste for mathematical formulae may wish to go directly to section 3.2, where some implications of MTC are discussed. The reader interested in

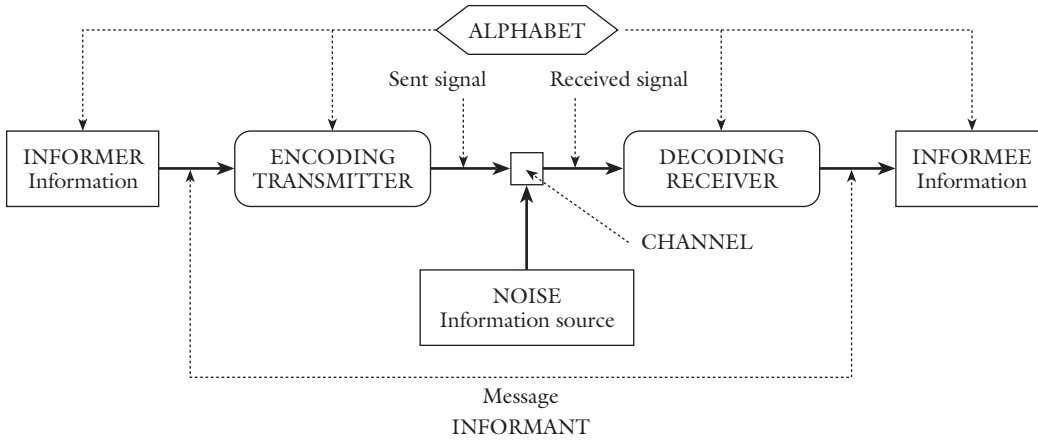

Figure 4.1: Communication model (adapted from Shannon 1948, 1998)

knowing more can start by reading Weaver 1949 and Shannon 1993b, then Schneider 2000, Pierce 1980, and Jones 1979, and finally Cover & Thomas 1991.

3.1 The quantification of raw information

MTC has its origin in the field of electrical communication, as the study of communication limits. It develops a quantitative approach to information as a means to answer two fundamental problems: the ultimate level of data compression and the ultimate rate of data transmission. The two solutions are the entropy H in equation (4.9) (see below) and the channel capacity C. The rest of this section illustrates how to get from the problems to the solutions.

Imagine a very boring device that can produce only one symbol, like Poe's raven, who can answer only "nevermore." This is called a *unary device*. Even at this elementary level, Shannon's simple model of communication applies (see figure 4.1). The raven is the *informer*, we are the *informee*, "nevermore" is the *message* (the informant), there is a coding and decoding procedure through a language (English), a *channel of communication*, and some possible *noise*.

Informer and informee share the same background knowledge about the collection of usable symbols (the *alphabet*). Given this *a priori* knowledge, it is obvious that a unary device produces zero amount of information. Simplifying, we already know the outcome so our ignorance cannot be decreased. Whatever the informational state of the system, asking appropriate questions to the raven does not make any difference. Note that a unary source answers every question all the time with only one symbol, not with silence or symbol, since silence counts as a signal, as we saw in section 2.1. It follows that a completely silent source also qualifies as a unary source.

Consider now a binary device that can produce two symbols, like a fair coin A with its two equiprobable symbols $\{h, t\}$; or, as Matthew 5:37 suggests, "Let your communication be Yea, yea; Nay, nay: for whatsoever is more than these cometh of evil." Before the coin is tossed, the informee (for example a computer) is in a state of *data deficit* greater than zero: the informee does not "know" which symbol the device will actually produce. Shannon used the technical term "uncertainty" to refer to data deficit. In a nonmathematical context this is a misleading term because of its strongly semantic connotations, especially from a Cartesian perspective. Recall that the informee can be a very simple machine, and psychological, mental, or doxastic states are clearly irrelevant. Once the coin has been tossed, the system produces an amount of raw information that is a function of the possible outputs, in this case 2 equiprobable symbols, and equal to the data deficit that it removes.

Table 4.1: Examples of communication devices and their informational power

Device	*Alphabet*	*Bits of information per symbol*
Poe's raven (unary)	1 symbol	log(1) = 0
1 coin (binary)	2 equiprobable symbols	log(2) = 1
2 coins	4 equiprobable symbols	log(4) = 2
1 die	6 equiprobable symbols	log(6) = 2.58
3 coins	8 equiprobable symbols	log(8) = 3

Let us build a slightly more complex system, made of two fair coins *A* and *B*. The *AB* system can produce 4 ordered outputs: <*h*, *h*>, <*h*, *t*>, <*t*, *h*>, <*t*, *t*>. It generates a data deficit of 4 units, each couple counting as a symbol in the source alphabet. In the *AB* system, the occurrence of each symbol <_, _> removes a higher data deficit than the occurrence of a symbol in the *A* system. In other words, each symbol provides more raw information. Adding an extra coin would produce a 8 units of data deficit, further increasing the amount of information carried by each symbol in the *ABC* system, and so on.

We are ready to generalize the examples. Call the number of possible symbols *N*. For $N = 1$, the amount of information produced by a unary device is 0. For $N = 2$, by producing an equiprobable symbol, the device delivers 1 unit of information. And for $N = 4$, by producing an equiprobable symbol the device delivers the sum of the amount of information provided by coin *A* plus the amount of information provided by coin *B*, that is, 2 units of information, although the total number of symbols is obtained by multiplying *A*'s symbols by *B*'s symbols. Now, our information measure should be a continuous and monotonic function of the probability of the symbols. The most efficient way of satisfying these requirements is by using the logarithm to the base 2 of the number of possible symbols (the logarithm to the base 2 of a number *n* is the power to which 2 must be raised to give the number *n*, for example $\log_2 8 = 3$, since $2^3 = 8$). Logarithms have the useful property of turning multiplication of symbols into addition of information units. By taking the logarithm to the base 2 (henceforth log simply means $\log_2$) we have the further advantage of expressing the units in bits. The base is partly a matter of convention, like using centimeters instead of inches, partly a matter of convenience, since it is useful when dealing with digital devices that use binary codes to represent data. Given an alphabet of *N* equiprobable symbols, we can rephrase some examples more precisely (table 4.1) by using equation (4.1):

$$\log_2(N) = \text{bits of information per symbol} \quad (4.1)$$

The basic idea is all in equation (4.1). Raw information can be quantified in terms of decrease in data deficit (Shannon's uncertainty). Unfortunately, real coins are always biased. To calculate how much information they produce one must rely on the frequency of the occurrences of symbols in a finite series of tosses, or on their probabilities, if the tosses are supposed to go on indefinitely. Compared to a fair coin, a slightly biased coin must produce less than 1 bit of information, but still more than 0. The raven produced no information at all because the occurrence of a string *S* of "nevermore" was not *informative* (not *surprising*, to use a more intuitive but psychologistic vocabulary), and that is because the *probability* of the occurrence of "nevermore" was maximum, so overly predictable. Likewise, the amount of raw information produced by the biased coin depends on the average *informativeness* (also known as average *surprisal*, another unfortunate term to refer to the average statistical rarity) of the string *S* of *h* and *t* produced by the coin. The average informativeness of the resulting string *S* depends on the *probability* of the occurrence of each symbol. The higher the frequency of a symbol in *S*, the less raw information is being produced by the

coin, up to the point when the coin is so biased to produce always the same symbol and stops being informative, behaving like the raven. So, to calculate the average informativeness of S we need to know how to calculate S and the informativeness of an i^{th} symbol in general. This requires understanding what the probability of an i^{th} symbol (P_i) to occur is.

The probability P_i of the i^{th} symbol can be "extracted" from equation (4.1), where it is embedded in $\log(N)$, a special case in which the symbols are equiprobable. Using some elementary properties of the logarithmic function we have:

$$\log(N) = -\log(N^{-1}) = -\log\left(\frac{1}{N}\right) = -\log(P) \tag{4.2}$$

The value of $1/N = P$ can range from 0 to 1. If the raven is our source, the probability of "good morning" is 0. In the case of the coin, $P(h) + P(t) = 1$, no matter how biased the coin is. Probability is like a cake that gets sliced more and more thinly depending on the number of guests, but never grows beyond its original size. More formally:

$$\sum_{i=1}^{N} P_i = 1 \tag{4.3}$$

The sigma notation simply means that if we add all probabilities values from $i = 1$ to $i = N$ the sum is equal to 1.

We can now be precise about the raven: "nevermore" is not informative at all because $P_{\text{nevermore}} = 1$. Clearly, the lower the probability of occurrence of a symbol, the higher is the informativeness of its actual occurrence. The informativeness u of an i^{th} symbol can be expressed by analogy with $-\log(P)$ in equation (4.2):

$$u_i = -\log(P_i) \tag{4.4}$$

Next, we need to calculate the length of a general string S. Suppose that the biased coin, tossed 10 times, produces the string: <h, h, t, h, h, t, t, h, h, t>. The (length of the) string S (in our case equal to 10) is equal to the number of times the h type of symbol occurs added to the numbers of times the t type of symbol occurs. Generalizing for i types of symbols:

$$S = \sum_{i=1}^{N} S_i \tag{4.5}$$

Putting together equations (4.4) and (4.5) we see that the average informativeness for a string of S symbols is the sum of the informativeness of each symbol divided by the sum of all symbols:

$$\frac{\sum_{i=1}^{N} S_i u_i}{\sum_{i=1}^{N} S_i} \tag{4.6}$$

Formula (4.6) can be simplified thus:

$$\sum_{i=1}^{N} \frac{S_i}{S} u_i \tag{4.7}$$

Now S_i/S is the frequency with which the i^{th} symbol occurs in S when S is finite. If the length of S is left undetermined (as long as one wishes), then the frequency of the i^{th} symbol becomes its probability P_i. So, further generalizing formula (4.7) we have:

$$\sum_{i=1}^{N} P_i u_i \tag{4.8}$$

Finally, by using equation (4.4) we can substitute for u_i and obtain

$$H = -\sum_{i=1}^{N} P_i \log P_i \text{ (bits per symbol)} \tag{4.9}$$

Equation (4.9) is Shannon's formula for H = uncertainty, which we have called *data deficit* (actually, Shannon's original formula includes a positive constant K which amounts to a choice of a unit of measure, bits in our case; apparently, Shannon used the letter H because of R. V. L. Hartley's previous work). Equation (4.9) indicates that the quantity of raw information produced by a device corresponds to the amount of

data deficit erased. It is a function of the average informativeness of the (potentially unlimited) string of symbols produced by the device. It is easy to prove that, if symbols are equiprobable, (4.9) reduces to (4.1) and that the highest quantity of raw information is produced by a system whose symbols are equiprobable (compare the fair coin to the biased one).

To arrive at (4.9) we have used some very simple examples: a raven and a handful of coins. Things in life are far more complex. For example, we have assumed that the strings of symbols are *ergodic*: the probability distribution for the occurrences of each symbol is assumed to be stable through time and independently of the selection of a certain string. Our raven and coins are *discrete* and *zero-memory sources*. The successive symbols they produce are statistically independent. But in real life occurrences of symbols are often interdependent. Sources can be non-ergodic and have a memory. Symbols can be continuous, and the occurrence of one symbol may depend upon a finite number *n* of preceding symbols, in which case the string is known as a Markov chain and the source an *n*th order Markov source. Consider for example the probability of being sent an "e" before or after having received the string "welcom." And consider the same example through time, in the case of a child learning how to spell English words. In brief, MTC develops the previous analysis to cover a whole variety of more complex cases. We shall stop here, however, because in the rest of this section we need to concentrate on other central aspects of MTC.

The quantitative approach just sketched plays a fundamental role in coding theory (hence in cryptography) and in data storage and transmission techniques. Recall that MTC is primarily a study of the properties of a channel of communication and of codes that can efficiently encipher data into recordable and transmittable signals. Since data can be distributed either in terms of here/there or now/then, diachronic communication and synchronic analysis of a memory can be based on the same principles and concepts (our coin becomes a bi-stable circuit or flip-flop, for example), two of which are so important to deserve a brief explanation: *redundancy* and *noise*.

Consider our *AB* system. Each symbol occurs with 0.25 probability. A simple way of encoding its symbols is to associate each of them with two digits:

$\langle h, h\rangle = 00$
$\langle h, t\rangle = 01$
$\langle t, h\rangle = 10$
$\langle t, t\rangle = 11$

Call this Code 1. In Code 1 a message conveys 2 bits of information, as expected. Do not confuse *bits* as *bi*-nary uni*ts* of information (recall that we decided to use $\log_2$ also as a matter of convenience) with *bits* as *bi*-nary digi*ts*, which is what a 2-symbols system like a CD-ROM uses to encode a message. Suppose now that the *AB* system is biased, and that the four symbols occur with the following probabilities:

$\langle h, h\rangle = 0.5$
$\langle h, t\rangle = 0.25$
$\langle t, h\rangle = 0.125$
$\langle t, t\rangle = 0.125$

This system produces less information, so by using Code 1 we would be wasting resources. A more efficient Code 2 should take into account the symbols' probabilities, with the following outcomes:

$\langle h, h\rangle = 0$	0.5×1 binary digit = .5
$\langle h, t\rangle = 10$	0.25×2 binary digits = .5
$\langle t, h\rangle = 110$	0.125×3 binary digits = .375
$\langle t, t\rangle = 111$	0.125×3 binary digits = .375

In Code 2, known as Fano Code, a message conveys 1.75 bits of information. One can prove that, given that probability distribution, no other coding system will do better than Fano Code. On the other hand, in real life a good codification is also modestly redundant. *Redundancy* refers to the difference between the physical representation of a message and the mathematical representation of the same message that uses no more bits than necessary. *Compression* procedures work by reducing data redundancy, but redundancy is not always a bad thing, for it can help to counteract *equivocation* (data sent but never received) and *noise* (received but unwanted data). A message + noise contains more data than the original message by itself, but the aim of a communication

process is *fidelity*, the accurate transfer of the original message from sender to receiver, not data increase. We are more likely to reconstruct a message correctly at the end of the transmission if some degree of redundancy counterbalances the inevitable noise and equivocation introduced by the physical process of communication and the environment. Noise extends the informee's freedom of choice in selecting a message, but it is an undesirable freedom and some redundancy can help to limit it. That is why, in a crowded pub, you shout your orders twice and add some gestures.

We are now ready to understand Shannon's two fundamental theorems. Suppose the 2-coins biased system produces the following message: <*t*, *h*> <*h*, *h*> <*t*, *t*> <*h*, *t*> <*h*, *t*>. Using Fano Code we obtain: 11001111010. The next step is to send this string through a channel. Channels have different transmission rates (C), calculated in terms of bits per second (bps). Shannon's fundamental theorem of the noiseless channel states that

> Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate of $C/H - \varepsilon$ symbols per second over the channel where ε is arbitrarily small. It is not possible to transmit at an average rate greater than C/H. (Shannon & Weaver 1998: 59)

In other words, if you devise a good code you can transmit symbols over a noiseless channel at an average rate as close to C/H as one may wish but, no matter how clever the coding is, that average can never exceed C/H. We have already seen that the task is made more difficult by the inevitable presence of noise. However, the fundamental theorem for a discrete channel with noise comes to our rescue:

> Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \varepsilon$ where ε is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$. (Shannon & Weaver 1998: 71)

Roughly, if the channel can transmit as much or more information than the source can produce, then one can devise an efficient way to code and transmit messages with as small an error probability as desired. These two fundamental theorems are among Shannon's greatest achievements. And with our message finally sent, we may close this section.

3.2 Some conceptual implications of MTC

For the mathematical theory of communication (MTC), information is only a selection of one symbol from a set of possible symbols, so a simple way of grasping how MTC quantifies raw information is by considering the number of yes/no questions required to guess what the source is communicating. One question is sufficient to guess the output of a fair coin, which therefore is said to produce 1 bit of information. A 2-fair-coins system produces 4 ordered outputs (<*h*, *h*>, <*h*, *t*>, <*t*, *h*>, <*t*, *t*>) and therefore requires two questions, each output containing 2 bits of information, and so on. This erotetic analysis clarifies two important points.

First, MTC is not a theory of information in the ordinary sense of the word. The expression "raw information" has been used to stress the fact that in MTC information has an entirely technical meaning. Consider some examples. Two equiprobable "yes" contain the same quantity of raw information, no matter whether their corresponding questions are "would you like some tea?" or "would you marry me?" If we knew that a device could send us with equal probabilities either the movie *Fahrenheit 451* or this whole *Guide*, by receiving one or the other we would receive many bytes of data but only one bit of raw information. On June 1, 1944, the BBC broadcasted a line from Verlaine's *Song of Autumn*: "Les sanglots longs des violons de

Autumne." The message contained almost 1 bit of information, an increasingly likely "yes" to the question whether the D-Day invasion was imminent. The BBC then broadcasted the second line "Blessent mon coeur d'une longueur monotone." Another almost meaningless string of letters, but almost another bit of information, since it was the other long-expected "yes" to the question whether the invasion was to take place immediately. German intelligence knew about the code, intercepted those messages, and even notified Berlin, but the high command failed to alert the Seventh Army Corps stationed in Normandy. Hitler had all the information in Shannon's sense of the word, but failed to understand the real meaning and importance of those two small bits of data. As for ourselves, we were not surprised to conclude that the maximum amount of raw information is produced by a text where each character is equally distributed, that is by a perfectly random sequence.

Second, since MTC is a theory of information without meaning, and information minus meaning = data, *mathematical theory of data communication* is a far more appropriate description than *information theory*. In section 2.1, we saw that information as semantic content can also be described erotetically as *data + queries*. Imagine a piece of information such as "the earth has only one moon." It is easy to polarize almost all its semantic content by transforming it into a query + binary answer: "does the earth have only one moon? + yes." Subtract the "yes" and you are left with virtually all the semantic content, fully de-alethicized (the query is neither true nor false). The datum "yes" works as a key to unlock the information contained in the query. MTC studies the codification and transmission of raw information by treating it as data keys, as the amount of details in a signal or message or memory space necessary to unlock the informee's knowledge. As Weaver (1949: 12) remarked, "the word information relates not so much to what you do say, as to what you could say. MTC deals with the carriers of information, symbols and signals, not with information itself. That is, information is the measure of your freedom of choice when you select a message."

Since MTC deals not with information itself but with the carriers of information, that is, messages constituted by uninterpreted symbols encoded in well-formed strings of signals, it is commonly described as a study of information at the *syntactic* level. MTC can be successfully applied in ICT (information and communication technologies) because computers are syntactical devices. What remains to be clarified is how H in equation (4.9) should be interpreted.

Assuming the ideal case of a noiseless channel of communication, H is a measure of three equivalent quantities:

a) the average amount of raw information per symbol produced by the informer, or
b) the corresponding average amount of data deficit (Shannon's uncertainty) that the informee has before the inspection of the output of the informer, or
c) the corresponding informational potentiality of the same source, that is, its *informational entropy*.

H can equally indicate (a) or (b) because, by selecting a particular alphabet, the informer automatically creates a data deficit (uncertainty) in the informee, which then can be satisfied (resolved) in various degrees by the *informant*. Recall the erotetic game. If you use a single fair coin, I immediately find myself in a 1-bit deficit predicament. Use two fair coins and my deficit doubles, but use the raven, and my deficit becomes null. My empty glass is an exact measure of your capacity to fill it. Of course, it makes sense to talk of raw information as quantified by H only if one can specify the probability distribution.

Regarding (c), MTC treats raw information like a physical quantity, such as mass or energy, and the closeness between equation (4.9) and the formulation of the concept of entropy in statistical mechanics was already discussed by Shannon. The informational and the thermodynamic concepts of entropy are related through the concepts of probability and *randomness* ("randomness" is better than "disorder" since the former is a syntactical concept whereas the latter has a strongly semantic value), entropy being a measure of the amount of "mixed-up-ness" in processes and systems bearing energy or information. Entropy can also be seen as an indicator of reversibility: if there is no change of entropy then the process is

reversible. A highly structured, perfectly organized message contains a lower degree of entropy or randomness, less raw information, and causes a smaller data deficit – consider the raven. The higher the potential randomness of the symbols in the alphabet, the more bits of information can be produced by the device. Entropy assumes its maximum value in the extreme case of uniform distribution. Which is to say that a glass of water with a cube of ice contains less entropy than the glass of water once the cube has melted, and a biased coin has less entropy than a fair coin. In thermodynamics, we know that the greater the entropy, the less available the energy. This means that high entropy corresponds to high energy deficit, but so does entropy in MTC: higher values of H correspond to higher quantities of data deficit.

4 Some Philosophical Approaches to Semantic Information

The mathematical theory of communication approaches information as a physical phenomenon. Its central question is whether and how much uninterpreted data can be encoded and transmitted efficiently by means of a given alphabet and through a given channel. MTC is not interested in the meaning, aboutness, relevance, usefulness, or interpretation of information, but only in the level of detail and frequency in the uninterpreted data, being these symbols, signals or messages. On the other hand, philosophical approaches seek to give an account of information as semantic content, investigating questions like "how can something count as information? and why?," "how can something carry information about something else?," "how can semantic information being generated and flow?," "how is information related to error, truth and knowledge?," "when is information useful?" Philosophers usually adopt a propositional orientation and an epistemic outlook, endorsing, often implicitly, the prevalence of the factual (they analyze examples like "The Bodleian library is in Oxford"). How relevant is MTC to similar analyses?

In the past, some research programs tried to elaborate information theories *alternative* to MTC, with the aim of incorporating the semantic dimension. Donald M. MacKay (1969) proposed a quantitative theory of qualitative information that has interesting connections with situation logic (see below), whereas Doede Nauta (1972) developed a semiotic-cybernetic approach. Nowadays, few philosophers follow these lines of research. The majority agree that MTC provides a rigorous constraint to any further theorizing on all the semantic and pragmatic aspects of information. The disagreement concerns the crucial issue of the *strength* of the constraint. At one extreme of the spectrum, a theory of semantic information is supposed to be *very strongly* constrained, perhaps even overdetermined, by MTC, somewhat as mechanical engineering is by Newtonian physics. Weaver's interpretation of Shannon's work is a typical example. At the other extreme, a theory is supposed to be *only weakly* constrained, perhaps even completely underdetermined, by MTC, somewhat as tennis is constrained by Newtonian physics, that is, in the most uninteresting, inconsequential, and hence disregardable sense (see for example Sloman 1978 and Thagard 1990). The emergence of MTC in the 1950s generated earlier philosophical enthusiasm that has gradually cooled down through the decades. Historically, philosophical theories of semantic information have moved from "very strongly constrained" to "only weakly constrained," becoming increasingly autonomous from MTC (for a review, see Floridi forthcoming *b*).

Popper (1935) is often credited as the first philosopher to have advocated the inverse relation between the probability of p and the amount of semantic information carried by p. However, systematic attempts to develop a formal calculus were made only after Shannon's breakthrough. MTC defines information in terms of probability space distribution. Along similar lines, the *probabilistic approach* to semantic information defines the semantic information in p in terms of logical probability space and the inverse relation between information and the probability of p. This approach was initially suggested by Bar-Hillel and Carnap (Bar-Hillel & Carnap 1953, Bar-Hillel 1964) and further developed by Hintikka (especially Hintikka & Suppes 1970) and Dretske 1981 (on Dretske's approach see also Chapters 16 and 17, on MEANING and on

KNOWLEDGE). The details are complex but the original idea is simple. The semantic content (CONT) in p is measured as the complement of the a priori probability of p:

$$\text{CONT}(p) = 1 - P(p) \tag{4.10}$$

CONT does not satisfy the two requirements of additivity and conditionalization, which are satisfied by another measure, the informativeness (INF) of p, which is calculated, following equations (4.9) and (4.10), as the reciprocal of $P(p)$, expressed in bits, where $P(p) = 1 - \text{CONT}(p)$:

$$\text{INF}(p) = \log \frac{1}{1 - \text{CONT}} = -\log P(p) \tag{4.11}$$

Things are complicated by the fact that the concept of probability employed in equations (4.10) and (4.11) is subject to different interpretations. In Bar-Hillel and Carnap the probability distribution is the outcome of a logical construction of atomic statements according to a chosen formal language. This introduces a problematic reliance on a strict correspondence between observational and formal language. In Dretske, the solution is to make probability values refer to states of affairs (s) of the world observed:

$$I(s) = -\log P(s) \tag{4.12}$$

The *modal approach* modifies the probabilistic approach by defining semantic information in terms of modal space and in/consistency. The information conveyed by p becomes the set of all possible worlds or (more cautiously) the set of all the descriptions of the relevant possible states of the universe that are excluded by p. The *systemic approach*, developed especially in situation logic (Barwise & Perry 1983, Israel & Perry 1990, Devlin 1991; Barwise & Seligman 1997 provide a foundation for a general theory of information flow) also defines information in terms of states space and consistency. However, it is less ontologically demanding than the modal approach, since it assumes a clearly limited domain of application, and it is compatible with Dretske's probabilistic approach, although it does not require a probability measure on sets of states. The informational content of p is not determined *a priori*, through a calculus of possible states allowed by a representational language, but in terms of factual content that p carries with respect to a given situation. Information tracks possible transitions in a system's states space under normal conditions. Both Dretske and situation theories require some presence of information already immanent in the environment (*environmental information*), as nomic regularities or constraints. This "semantic externalism" can be controversial both epistemologically and ontologically. Finally, the *inferential approach* defines information in terms of entailment space: information depends on valid inference relative to a person's theory or epistemic state.

Most approaches close to MTC assume the principle of *alethic neutrality*, and run into the difficulties I outlined in section 2.2 (Dretske and Barwise are important exceptions; Devlin rejects truthfulness as a necessary condition). As a result, the *semantic approach* (Floridi forthcoming *a* and *c*) adopts SDI and defines factual information in terms of data space.

Suppose there will be exactly three guests for dinner tonight. This is our situation w. Imagine that you are told that

T) there may or may not be some guests for dinner tonight; or
V) there will be some guests tonight; or
P) there will be three guests tonight.

The *degree of informativeness* of T is zero because, as a tautology, T applies both to w and to $\neg w$. V performs better, and P has the maximum degree of informativeness because, as a fully accurate, precise, and contingent truth, it "zeros in" on its target w. Generalizing, the more distant a true σ is from its target w, the larger is the number of situations to which it applies, the lower its degree of informativeness becomes. A tautology is a true σ that is most "distant" from the world. Let us use the letter ϑ to refer to the distance between a true σ and w. Using the more precise vocabulary of situation logic, ϑ indicates the degree of support offered by w to σ. We can now map on the x axis the values of ϑ given a specific σ and a corresponding target w. In our

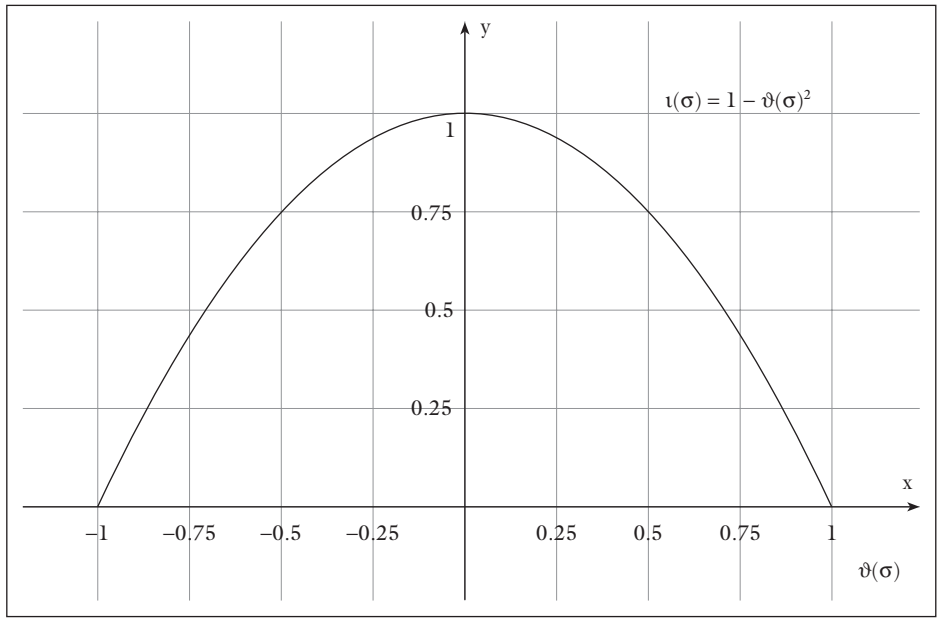

Figure 4.2: degree of informativeness (ι) of σ

example, we know that ϑ(T) = 1 and ϑ(P) = 0. For the sake of simplicity, let us assume that ϑ(V) = 0.25 (see Floridi forthcoming *c* on how to calculate ϑ values). We now need a formula to calculate the *degree of informativeness* ι of σ in relation to ϑ(σ). It can be shown that the most elegant solution is provided by the complement of the square value of ϑ(σ), that is $y = 1 - x^2$. Using the symbols jut introduced we have:

$$\iota(\sigma) = 1 - \vartheta(\sigma)^2 \qquad (4.13)$$

Figure 4.2 shows the graph generated by equation (4.13) when we also include negative values of distance for false σ (ϑ ranges from –1 = contradiction to 1 = tautology).

If σ has a very high degree of informativeness ι (very low ϑ) we want to be able to say that it contains a large quantity of semantic information and, vice versa, the lower the degree of informativeness of σ is, the smaller the quantity of semantic information conveyed by σ should be. To calculate the quantity of semantic information contained in σ relative to ι(σ) we need to calculate the area delimited by equation (4.13), that is, the definite integral of the function ι(σ) on the interval [0, 1]. As we know, the maximum quantity of semantic information (call it α) is carried by P, whose ϑ = 0. This is equivalent to the whole area delimited by the curve. Generalizing to σ we have:

$$\int_0^1 \iota\,(\sigma)\,dx = \alpha = \frac{2}{3} \qquad (4.14)$$

Figure 4.3 shows the graph generated by equation (4.14). The shaded area is the maximum amount of semantic information α carried by σ.

Consider now V, "there will be some guests tonight." V can be analyzed as a (reasonably finite) string of disjunctions, that is V = ["there will be one guest tonight" or "there will be two guests tonight" or . . . "there will be *n* guests tonight"], where *n* is the reasonable limit we wish to consider (things are more complex than this, but here we only need to grasp the general principle). Only one of the descriptions in V will be fully accurate. This means that V also contains some (perhaps much) information that is simply irrelevant or redundant. We shall refer

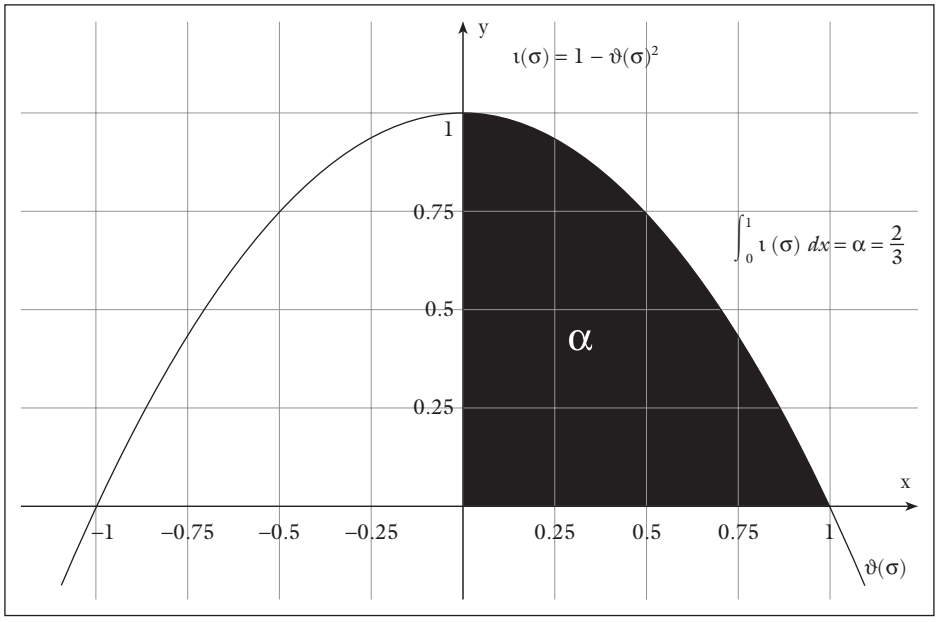

Figure 4.3: maximum amount of semantic information α carried by σ

to this "informational waste" in V as vacuous information in V. The amount of vacuous information (call it β) in V is also a function of the distance ϑ of V from *w*, or more generally

$$\int_0^{\vartheta} \iota(\sigma)\, dx = \beta \qquad (4.16)$$

Since ϑ(V) = 0.25, we have

$$\int_0^{0.25} \iota(\mathrm{V})\, dx = 0.24479 \qquad (4.17)$$

Figure 4.4 shows the graph generated by equation (4.17). The shaded area is the amount of vacuous information β in V. Clearly, the amount of semantic information in V is simply the difference between α (the maximum amount of information that can be carried in principle by σ) and β (the amount of vacuous information actually carried by σ), that is, the clear area in the graph of figure 4.4. More generally, the amount of semantic information γ in σ is:

$$\gamma(\sigma) = (\alpha - \beta) \qquad (4.18)$$

Note the similarity between 4.14 and 4.16. When ϑ(σ) = 1, that is, when the distance between σ and *w* is maximum, then α = β and γ(σ) = 0. This is what happens when we consider T. T is so distant from *w* as to contain only vacuous information. In other words, T contains as much vacuous information as P contains relevant information.

A final comment, before closing this section. Each of the previous extentionalist approaches can be given an intentionalist interpretation by considering the relevant space as a doxastic space, in which information is seen as a reduction in the degree of personal uncertainty given a state of knowledge of the informee.

5 Conclusion

In this chapter, we have been able to visit only a few interesting places. The connoisseur might be disappointed and the supporter of some local interests appalled. To try to appease both and to whet the appetite of the beginner here is a list of

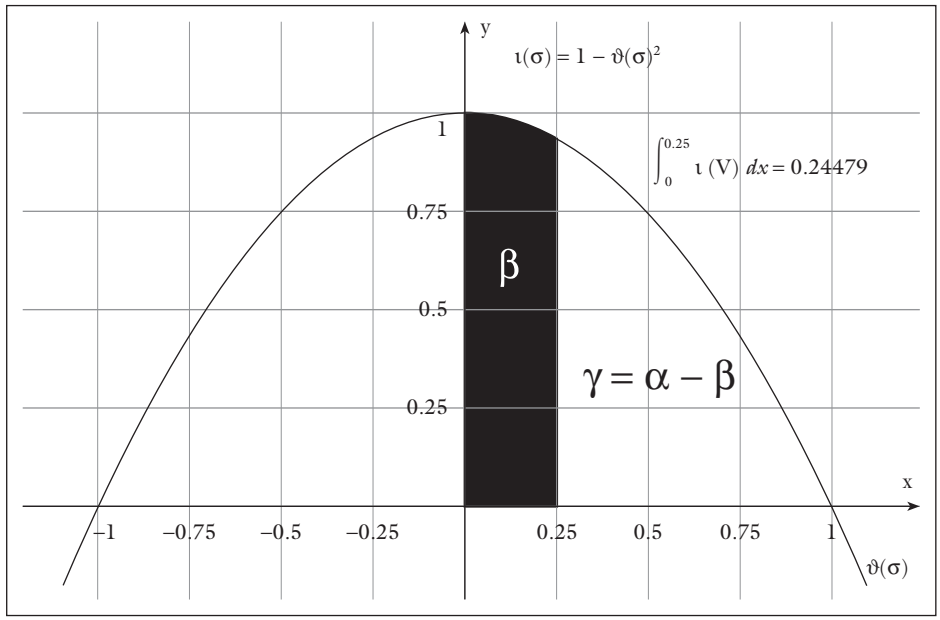

Figure 4.4: Amount of semantic information γ carried by σ

some very important concepts of information that have not been discussed:

informational complexity (Kolmogorov and Chaitin, among others), a measure of the complexity of a string of data defined in terms of the length of the shortest binary program required to compute that string. Note that Shannon's H can be considered a special case of Kolmogorov complexity K, since $H \approx K$ if the sequence is drawn at random from a probability distribution with entropy = H;

instructional information (imagine a recipe, an algorithm, or an order), a crucial concept in fields like computer science, genetics, biochemistry, neuroscience, cognitive science, and AI (see Chapters 1 and 2, on Computation and on Complexity);

pragmatic information, central in any theory addressing the question of how much information a certain informant carries for an informee in a given doxastic state and within a specific informational environment. This includes *useful information*, a key concept in economics, information management theory, and decision theory, where characteristics such as relevance, timeliness, updatedness, cost, significance, and so forth are crucial (see Chapter 22, on Game Theory);

valuable information in ethical contexts (see Chapter 5, on Computer Ethics, and Floridi 2003);

environmental information, that is, the possible location and nature of information in the world (Dretske 1981 and see Chapters 11–13, on Ontology, on Virtual Reality, and on The Physics of Information, respectively);

physical information and the relation between being and information (see Leff & Rex 1990 and again Chapters 11–13);

biological information (see Chapter 15, on Artificial Life). The biologically minded reader will notice that the 4 symbols in the AB system we built in section 3.1 could be adenine, guanine, cytosine, and thymine, the four bases whose order in the molecular chain of DNA or RNA codes genetic information.

The nature of these and other information concepts, the analysis of their interrelations and

of their possible dependence on MTC, and the investigation of their usefulness and influence in the discussion of philosophical problems are some of the crucial issues that a philosophy of information needs to address. There is clearly plenty of very interesting and important work to do.

Acknowledgments

This chapter is based on Floridi 2003 and forthcoming *a*. I am very grateful to Mark Bedau, John Collier, Phil Fraundorf, Ken Herold, James Fetzer, and Kia Nobre for their very valuable comments on earlier drafts.

Some Web Resources

There are many useful resources freely available on the web, of which the following have been used in writing this chapter:

Feldman D., *A Brief Tutorial on Information Theory, Excess Entropy and Statistical Complexity*: <http://hornacek.coa.edu/dave/Tutorial/index.html>

Fraundorf P., *Information-Physics on the Web*: <http://newton.umsl.edu/infophys/infophys.html>

Introduction to Information Theory, by Lucent Technologies Bell Labs Innovation: <http://www.lucent.com/minds/infotheory/>

MacKay J. C., *A Short Course in Information Theory*: <http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html>

Shannon C. E. 1948. *A Mathematical Theory of Communication*: <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html> [The classic text on the mathematical theory of information, graduate level.]

Schneider T. 2000. *Information Theory Primer – With an Appendix on Logarithms*: <http://www-lmmb.ncifcrf.gov/~toms/paper/primer/index.html> [A very clear and accessible introduction, undergraduate level.]

UTI, the Unified Theory of Information website, contains documents, and links about the development of UTI**:** <http://kaneda.iguw.tuwien.ac.at/uti/uti4/index.html>

References

Bar-Hillel, Y. 1964. *Language and Information*. Reading, MA and London: Addison Wesley. [Collection of influential essays on semantic information, graduate level.]

—— and Carnap, R. 1953. "An outline of a theory of semantic information," repr. in Bar-Hillel 1964: 221–74. [One of the first and most influential attempts to develop a quantitative analysis of semantic information, graduate level.]

Barwise, J. and Perry, J. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press. [Influential text in situation logic, graduate level.]

—— and Seligman, J. 1997. *Information Flow: The Logic of Distributed Systems*. Cambridge: Cambridge University Press. [Innovative approach to information flow, graduate level, but the modular structure contains some very accessible chapters.]

Bateson, G. 1973. *Steps to an Ecology of Mind*. Frogmore, St. Albans: Paladin. [The beginning of the ecological approach to mind and information, undergraduate level.]

Braman, S. 1989. "Defining information." *Telecommunications Policy* 13: 233–42.

Chalmers, D. J. 1997. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press. [Accessible to undergraduates.]

Cherry, C. 1978. *On Human Communication*, 3rd ed. Cambridge, MA: MIT Press. [Clear and accessible introduction to communication theory, old but still valuable.]

Colburn, T. R. 2000. "Information, thought, and knowledge." *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics* (Orlando, FL, July 23–6, 2000), vol. 10: 467–71. [Analyzes the standard definition of knowledge as justified true belief from an informational perspective, very accessible.]

Cover, T. and Thomas, J. A. 1991. *Elements of Information Theory*. New York: Chichester, Wiley. [Standard textbook in the field, requires a solid mathematical background, graduate level only, see Jones 1979 for a more accessible text, or Pierce 1980.]

Deutsch, D. 1985. "Quantum theory, the Church–Turing Principle and the Universal Quantum Computer." *Proceedings of the Royal Society* 400: 97–117. [Information and computation in quantum computing, requires a solid mathematical background, graduate level only.]

——. 1997. *The Fabric of Reality*. London: Penguin. [On the ontological implications of quantum physics, advanced undergraduate level.]

Devlin, K. 1991. *Logic and Information*. Cambridge: Cambridge University Press. [Reviews and improves on situation logic, undergraduate level.]

Di Vincenzo, D. P. and Loss, D. 1998. "Quantum information is physical." *Superlattices and Microstructures* 23: 419–32; special issue on the occasion of Rolf Landauer's 70th birthday; also available at <http://xxx.lanl.gov/abs/cond-mat/9710259>. [Reviews the debate on the physical aspects of information, graduate level.]

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press, rep. Stanford: CSLI, 1999. [Classic informational analysis of knowledge, advanced undergraduate level.]

Fetzer, J. H. (forthcoming). "Information misinformation, and disinformation." Forthcoming in *Minds and Machines*. [Criticizes Floridi's view that information encapsulates truth and develops an alternative account; accessible.]

Floridi, L. 1999. *Philosophy and Computing – An Introduction*. London and New York: Routledge. [Textbook that complements this *Guide*, elementary undergraduate level.]

——. 2003. "On the intrinsic value of information objects and the infosphere.". *Ethics and Information Technology* 4, no. 4: 287–304. Preprint available at <http://www.wolfson.ox.ac.uk/~floridi/papers.htm>. [Develops an ethical approach to information environments, advanced undergraduate level.>

——. (forthcoming *a*). "Is semantic information meaningful data?" Forthcoming in *Philosophy and Phenomenological Research*. Preprint available at <http://www.wolfson.ox.ac.uk/~floridi/papers.htm>. [Defines semantic information as well-formed, meaningful, and truthful data; graduate level.]

——. (forthcoming *b*). "Information, semantic conceptions of." *Stanford Encyclopedia of Philosophy*. [Reviews philosophical conceptions of semantic information, undergraduate level.]

——. (forthcoming *c*) "Outline of a theory of strongly semantic information." Forthcoming in *Minds and Machines*. Preprint available at <http://www.wolfson.ox.ac.uk/~floridi/papers.htm>. [Develops a truth-based approach to semantic information, graduate level.]

Fox, C. J. 1983. *Information and Misinformation – An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming*. Westport, CN: Greenwood Press. [Analysis of information based on information science, undergraduate level.]

Franklin, S. 1995. *Artificial Minds*. Cambridge, MA: MIT Press. [Undergraduate level.]

Frieden, B. R. 1998. *Physics from Fisher Information: A Unification*. Cambridge: Cambridge University Press. [Controversial attempt to provide an interpretation of physics in terms of information, requires a solid background in mathematics, graduate level only.]

Grice, P. 1989. *Studies in the Way of Words*. Cambridge MA: Harvard University Press. [Collection of Grice's influential works, advanced undergraduate level.]

Hanson, P., ed. 1990. *Information, Language and Cognition*. Vancouver: University of British Columbia Press. [Important collection of essays, most at graduate level.]

Hintikka, J. and Suppes, P. 1970. *Information and Inference*. Dordrecht: Reidel. [Important collection of philosophical essays on information theory, graduate level.]

Israel, D. and Perry, J. 1990. "What is information?" In Hanson 1990: 1–19. [Analyzes information on the basis of situation logic, graduate level.]

Jones, D. S. 1979. *Elementary Information Theory*. Oxford: Clarendon Press. [Brief textbook on information theory, less mathematical than Cover & Thomas 1991, but still more demanding than Pierce 1980.]

Landauer, R. 1987. "Computation: a fundamental physical view." *Physica Scripta* 35: 88–95. [Graduate level only.]

——. 1991. "Information is physical." *Physics Today* 44: 23–9. [Graduate level only.]

——. 1996. "The physical nature of information." *Physics Letter* A217: 188. [Graduate level only.]

—— and Bennett, C. H. 1985. "The fundamental physical limits of computation." *Scientific American* July: 48–56. [A more accessible presentation of the view that information requires a physical implementation, undergraduate level.]

Leff, H. S. and Rex, A. F. 1990. *Maxwell's Demon: Entropy, Information, and Computing*. Bristol:

Hilger. [Collection of essays on this classic problem, graduate level.]

Losee, R. M. 1997. "A discipline independent definition of information." *Journal of the American Society for Information Science* 48(3): 254–69. [Undergraduate level.]

Mabon, P. C. 1975. *Mission Communications: The Story of Bell Laboratories.* [Very readable account of the people and the discoveries that made information theory possible, undergraduate level.]

Machlup, F. 1983. "Semantic quirks in studies of information." In F. Machlup and U. Mansfield, eds., *The Study of Information: Interdisciplinary Messages.* New York: John Wiley, pp. 641–71. [Advanced undergraduate level.]

MacKay, D. M. 1969. *Information, Mechanism and Meaning.* Cambridge, MA: MIT Press. [Develops an alternative view of information to Shannon's, graduate level.]

Mingers, J. 1997. "The nature of information and its relationship to meaning." In R. L. Winder et al., eds., *Philosophical Aspects of Information Systems.* London: Taylor and Francis, pp. 73–84. [Analyzes information from a system theory perspective, advanced undergraduate level.]

NATO. 1974. Advanced Study Institute in Information Science, Champion, 1972. *Information Science: Search for Identity*, ed. A. Debons. New York: Marcel Dekker.

——. 1975. Advanced Study Institute in Information Science, Aberystwyth, 1974. *Perspectives in Information Science*, eds. A. Debons and W. J. Cameron. Leiden: Noordhoff.

——. 1983. Advanced Study Institute in Information Science, Crete, 1978. *Information Science in Action: Systems Design*, eds. A. Debons and A. G. Larson. Boston: Martinus Nijhoff.

Nauta, D. 1972. *The Meaning of Information.* The Hague: Mouton. [Reviews various analyses of information, advanced undergraduate level.]

Newell, A. and Simon, H. A. 1976. "Computer science as empirical inquiry: symbols and search." *Communications of the ACM*, 19 March: 113–26. [The classic paper presenting the Physical Symbol System Hypothesis in AI and cognitive science, graduate level.]

Pierce, J. R. 1980. *An Introduction to Information Theory: Symbols, Signals and Noise.* New York, Dover Publications. [Old but still very valuable introduction to information theory for the non-mathematician, undergraduate level.]

Popper, K. R. 1935. *Logik der Forschung: zur Erkenntnistheorie der modernen Naturwissenschaft.* Vienna: J. Springer; tr. *The Logic of Scientific Discovery.* London: Hutchinson, 1959. [Popper's classic text, graduate level.]

Schrader, A. 1984. "In search of a name: information science and its conceptual antecedents." *Library and Information Science Research* 6: 227–71. [Undergraduate level.]

Schneider, T. 2000. "Information theory primer – with an appendix on logarithms." Version 2.48, postscript version <ftp://ftp.ncifcrf.gov/pub/delila/primer.ps>, web version <http://www.lecb.ncifcrf.gov/~toms/paper/primer/>. [A very clear and simple introduction that can also be consulted for further clarification about the mathematics involved, undergraduate level.]

Shannon, C. E. 1993a. *Collected Papers*, eds. N. J. A. Sloane and A. D. Wyner. Los Alamos, CA: IEEE Computer Society Press. [Mostly graduate level only.]

——. 1993b. Article on "Information Theory," *Encyclopedia Britannica*, repr. in his *Collected Papers*, pp. 212–20. [A brief and accessible presentation of information theory by its founding father, undergraduate level.]

—— and Weaver, W. 1998 [first published 1948]. *The Mathematical Theory of Communication*, with a foreword by R. E. Blahut and B. Hajek. Urbana and Chicago, IL: University of Illinois Press. [The classic text in information theory, graduate level; Shannon's text is also available on the web – see above.]

Sloman A. 1978. *The Computer Revolution in Philosophy.* Atlantic Highlands: Humanities Press. [One of the earliest and most insightful discussions of the informational/computation turn in philosophy, most chapters undergraduate level.]

Steane, A. M. 1998. "Quantum computing." *Reports on Progress in Physics* 61: 117–73. [A review, graduate level, also available online at <http://xxx.lanl.gov/abs/quant-ph/9708022>.]

Thagard, P. R. 1990. "Comment: concepts of information." In Hanson 1990.

Weaver, W. 1949. "The mathematics of communication." *Scientific American* 181(1): 11–15. [Very accessible introduction to Shannon's theory, undergraduate level.]

Wellisch, H. 1972. "From information science to informatics." *Journal of Librarianship* 4: 157–87.

Wersig, G. and Neveling, U. 1975. "The phenomena of interest to information science." *Information Scientist* 9: 127–40.

Wheeler, J. A. 1990. "Information, physics, quantum: the search for links." In W. H. Zureck, ed., *Complexity, Entropy, and the Physics of Information.* Redwood City, CA: Addison Wesley. [Introduces the "It from Bit" hypothesis, graduate level.]

Wiener, N. 1954. *The Human Use of Human Beings: Cybernetics and Society*, 2nd ed.; reissued in 1989 with a new introduction by Steve J. Heims. London: Free Association. [A very early discussion of the ethical and social implications of the computer revolution, undergraduate level.]

——. 1961. *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, MA: MIT Press. [The foundation of cybernetics, graduate level.]

Part II

Computers in Society

Chapter 5

Computer Ethics

Deborah G. Johnson

1 Introduction

From the moment of their invention, computers have generated complex social, ethical, and value concerns. These concerns have been expressed in a variety of ways, from the science fiction stories of Isaac Asimov (1970) to a dense three-volume treatise on social theory by Manuel Castells (1996, 1997, 1998), and with much in between. Generally, the literature describes the social consequences of computing, speculates on the meaning of computation and information technology in human history, and creatively predicts the future path of development of computer technology and social institutions around it. A small, though steadily increasing, number of philosophers has focused specifically on the *ethical issues*.

As computer technology evolves and gets deployed in new ways, certain issues persist – issues of privacy, property rights, accountability, and social values. At the same time, seemingly new and unique issues emerge. The ethical issues can be organized in at least three different ways: according to the type of technology; according to the sector in which the technology is used; and according to ethical concepts or themes. In this chapter I will take the third approach. However, before doing so it will be useful to briefly describe the other two approaches.

The first is to organize the ethical issues by type of technology and its use. When computers were first invented, they were understood to be essentially sophisticated calculating machines, but they seemed to have the capacity to do that which was thought to be uniquely human – to reason and exhibit a high degree of rationality; hence, there was concern that computers threatened ideas about what it means to be human. In the shadow of the Second World War, concerns quickly turned to the use of computers by governments to centralize and concentrate power. These concerns accompanied the expanding use of computers for record-keeping and the exponential growth in the scale of databases, allowing the creation, maintenance, and manipulation of huge quantities of personal information. This was followed by the inception of software control systems and video games, raising issues of accountability–liability and property rights. This evolution of computer technology can be followed through to more recent developments including the internet, simulation and imaging technologies, and virtual reality systems. Each one of these developments was accompanied by conceptual and moral uncertainty. What will this or that development mean for the lives and values of human beings? What will it do to the relationship between government and citizen? Between employer and employee? Between businesses and consumers?

A second enlightening approach is to organize the issues according to the sector in which they occur. Ethical issues arise in real-world contexts, and computer-ethical issues arise in the contexts in which computers are used. Each context or sector has distinctive issues, and if we ignore this context we can miss important aspects of computer-ethical issues. For example, in dealing with privacy protection in general, we might miss the special importance of privacy protection for *medical records* where confidentiality is so essential to the doctor–patient relationship. Similarly, one might not fully understand the appropriate role for computers in education were one not sensitive to distinctive goals of education.

Both of these approaches – examining issues by types and uses of particular technologies, and sector by sector – are important and illuminating; however, they take us too far afield of the philosophical issues. The third approach – the approach to be taken in this chapter – is to emphasize ethical concepts and themes that persist across types of technology and sectors. Here the issues are sorted by their philosophical and ethical content. In this chapter I divide the issues into two broad categories: (1) metatheoretical and methodological issues, and (2) traditional and emerging issues.

2 Metatheoretical and Methodological Issues

Perhaps the deepest philosophical thinking on computer-ethical issues has been reflection on the field itself – its appropriate subject matter, its relationship to other fields, and its methodology. In a seminal piece entitled "What is Computer Ethics?" Moor (1985) recognized that when computers are first introduced into an environment, they make it possible for human beings (individuals and institutions) to do things they couldn't do before, and this creates *policy vacuums.* We do not have rules, policies, and conventions on how to behave with regard to the new possibilities. Should employers monitor employees to the extent possible with computer software? Should doctors perform surgery remotely? Should I make copies of proprietary software? Is there any harm in me taking on a pseudo-identity in an online chatroom? Should companies doing business online be allowed to sell the transaction-generated information they collect? These are examples of policy vacuums created by computer technology.

Moor's account of computer ethics has shaped the field of computer ethics with many computer ethicists understanding their task to be that of helping to fill policy vacuums. Indeed, one of the topics of interest in computer ethics is to understand this activity of filling policy vacuums. This will be addressed later on.

2.1 The connection between technology and ethics

While Moor's account of computer ethics remains influential, it leaves several questions unanswered. Hence, discussion and debate continue around the question of why there is or should be a field of computer ethics and what the focus of the field should be.

In one of the deeper analyses, Floridi (1999) argues for a metaphysical foundation for computer ethics. He provides an account of computer ethics in which information has status such that destroying information can itself be morally wrong. In my own work I have tried to establish the foundation of computer ethics in the non-obvious connection between technology and ethics (Johnson 2001). Why is technology of relevance to ethics? What difference can technology make to human action? To human affairs? To moral concepts or theories?

Two steps are involved in answering these questions. The first step involves fully recognizing something that Moor's account acknowledges, namely that technology often makes it possible for human beings to do what they could not do without it. Think of spaceships that take human beings to the moon; think of imaging technology that allows us to view internal organs; or think of computer viruses that wreak havoc on the internet.

Of course, it is not just that human beings can do what they couldn't do before. It is also that we can do the same sorts of things we did before, only in new ways. As a result of technology, we

can travel, work, keep records, be entertained, communicate, and engage in warfare *in new ways.* When we engage in these activities using computer technology, our actions have different properties, properties that may change the character of the activity or action-type. Consider the act of writing with various technologies. When I write with paper and pencil, the pencil moves over paper; when I write using a typewriter, levers and gears move; when I write using a computer, electronic impulses change configurations in microchips. So, the physical events that take place when I write are very different when I use computer technology.

Using action theory, the change can be characterized as a change in the possible act tokens of an act type. An act type is a kind of action (e.g. reading a book, walking) and an act token is a particular instance of an act type. An act token is an instance of the act type performed by a particular person, at a particular time, and in a particular place. For example, "Jan is, at this moment, playing chess with Jim in Room 200 of Thornton Hall on the campus of University of Virginia" is an act token of the act type "playing chess." When technology is involved in the performance of an act type, a new set of act tokens may become possible. It is now possible, for example, to "play chess" while sitting in front of a computer and not involving another human being. Instead of manually moving three-dimensional pieces, one presses keys on a keyboard or clicks on a mouse. Thus, when human beings perform actions with computers, new sets of tokens (of act types) become possible. Most important, the new act tokens have properties that are distinct from other tokens of the same act type.

Computer technology instruments human action in ways that turn very simple movements into very powerful actions. Consider hardly-visible finger movements on a keyboard. When the keyboard is connected to a computer and the computer is connected to the internet, and when the simple finger movements create and launch a computer virus, those simple finger movements can wreak havoc in the lives of thousands (even millions) of people. The technology has instrumented an action not possible without it. To be sure, individuals could wreak havoc on the lives of others before computer technology, but not in this way and perhaps not quite so easily. Computer technology is not unique among technologies in this respect; other technologies have turned simple movements of the body into powerful actions, e.g. dynamite, automobiles.

Recognizing the intimate connection between technology and human action is important for stopping the deflection of human responsibility in technology-instrumented activities, especially when something goes wrong. Hence, the hacker cannot avoid responsibility for launching a virus on grounds that he simply moved his fingers while sitting in his home. Technology does nothing independent of human initiative; though, of course, sometimes human beings cannot foresee what it is they are doing with technology.

Thus, the first step in understanding the connection between computer technology and ethics is to acknowledge how intimate the connection between (computer) technology and human action can be. The second step is to connect human action to ethics. This step may seem too obvious to be worthy of mention since ethics is often understood to be exclusively the domain of human action. Even so, computer technology changes the domain of human action; hence, it is worth asking whether these changes have moral significance. Does the involvement of computer technology – in a human situation – have moral significance? Does the *instrumentation* of human action affect the character of ethical issues, the nature of ethical theory, or ethical decision-making?

The involvement of computer technology has moral significance for several reasons. As mentioned earlier, technology creates new possibilities for human action and this means that human beings face ethical questions they never faced before. Should we develop biological weapons and risk a biological war? Should I give my organs for transplantation? In the case of computer technology, is it wrong to monitor keystrokes of employees who are using computers? To place cookies on computers when the computers are used to visit a website? To combine separate pieces of personal data into a single comprehensive portfolio of a person?

When technology changes the properties of tokens of an act type, the moral character of the

act type can change. In workplace monitoring, for example, while it is generally morally acceptable for employers to keep track of the work of employees, the creation of software that allows the employer to record and analyze every keystroke an employee makes raises the question in a new way. The rights of employers and employees have to be reconsidered in light of this new possibility. Or to use a different sort of example, when it comes to property rights in software, the notion of property and the stakes in owning and copying are significantly different when it comes to computer software because computer software has properties unlike that of anything else. Most notably, software can be replicated with no loss to the owner in terms of possession or usefulness (though, of course, there is a loss in the value of the software in the marketplace).

So, computers and ethics are connected insofar as computers make it possible for humans to do things they couldn't do before and to do things they could do before but in new ways. These changes often have moral significance.

2 Applied and Synthetic Ethics

To say that computer technology creates new tokens of an act type may lead some to categorize computer ethics as a branch of applied or practical ethics. Once a computer ethical issue is understood to involve familiar act types, it might be presumed, all that is necessary to resolve the issue is to use moral principles and theories that generally apply to the act type. For example, if the situation involves honesty in communicating information, simply follow the principle, "tell the truth," with all its special conditions and caveats. Or, if the situation involves producing some positive and negative effects, simply do the utilitarian calculation. This account of computer ethics is, however, as controversial as is the notion of "applied ethics" more generally.

For one thing, computer technology and the human situations arising around it are not always so easy to understand. As Moor has pointed out, often there are conceptual muddles (1985). What is software? What is a computer virus? How are we to conceptualize a search engine? A cookie? A virtual harm? In other words, computer ethicists do more than "apply" principles and theories; they do conceptual analysis. Moreover, the analysis of a computer-ethical issue often involves synthesis, synthesis that creates an understanding of both the technology and the ethical situation. A fascinating illustration of this is the case of a virtual rape (Dibbell 1993). Here a character in a multi-user virtual reality game rapes another character. Those participating in the game are outraged and consider the behavior of the real person controlling the virtual characters offensive and bad. The computer ethical issue involves figuring out what, if anything, wrong the real person controlling the virtual character has done. This involves understanding how the technology works, what the real person did, figuring out how to characterize the actions, and then recommending how the behavior should be viewed and responded to. Again, analysis of this kind involves more than simply "applying" principles and theories. It involves conceptual analysis and interpretation. Indeed, the synthetic analysis may have implications that reflect back on the meaning of, or our understanding of, familiar moral principles and theories.

To be sure, philosophical work in computer ethics often does involve drawing on and extending the work of well-known philosophers and making use of familiar moral concepts, principles, and theories. For example, computer ethical issues have frequently been framed in utilitarian, deontological, and social contract theory. Many scholars writing about the internet have drawn on the work of existentialist philosophers such as Søren Kierkegaard (Dreyfus 1999; Prosser & Ward 2000) and Gabriel Marcel (Anderson 2000). The work of Jürgen Habermas has been an important influence on scholars working on computer-mediated communication (Ess 1996). Recently van den Hoven (1999) has used Michael Walzer's "spheres of justice" to analyze the information society; Cohen (2000) and Introna (2001) have used Emmanuel Levinas to understand internet communication; Adams and Ofori-Amanfo (2000) have been connecting feminist ethics to computer ethics; and Grodzinsky (1999) has developed virtue theory to illuminate computer ethics.

Nevertheless, while computer ethicists often draw on, extend, and "apply" moral concepts and theories, computer ethics involves much more than this. Brey (2000) has recently argued for an approach that he labels "disclosive computer ethics." The applied ethics model, he notes, emphasizes controversial issues for which the ethical component is transparent. Brey argues that there are many nontransparent issues, issues that are not so readily recognized. Analysis must be done to "disclose" and make visible the values at stake in the design and use of computer technology. A salient example here is work by Introna and Nissenbaum (2000) on search engines. They show how the design of search engines is laden with value choices. In order to address those value choices explicitly, the values embedded in search engine design must be uncovered and disclosed. This may sound simple but in fact uncovering the values embedded in technology involves understanding how the technology works and how it affects human behavior and human values.

Setting aside what is the best account of computer ethics, it should be clear that a major concern of the field is to understand its domain, its methodology, its reason for being, and its relationship to other areas of ethical inquiry. As computer technology evolves and gets deployed in new ways, more and more ethical issues are likely to arise.

3 Traditional and Emerging Issues

"Information society" is the term often used (especially by economists and sociologists) to characterize societies in which human activity and social institutions have been significantly transformed by computer and information technology. Using this term, computer ethics can be thought of as the field that examines ethical issues distinctive to "an information society." Here I will focus on a subset of these issues, those having to do with professional ethics, privacy, cyber crime, virtual reality, and general characteristics of the internet.

3.1 Ethics for computer professionals

In an information society, a large number of individuals are educated for and employed in jobs that involve development, maintenance, buying and selling, and use of computer and information technology. Indeed, an information society is dependent on such individuals – dependent on their special knowledge and expertise and on their fulfilling correlative social responsibilities. Expertise in computing can be deployed recklessly or cautiously, used for good or ill, and the organization of information technology experts into occupations/professions is an important social means of managing that expertise in ways that serve human well-being.

An important philosophical issue here has to do with understanding and justifying the social responsibilities of computer experts. Recognizing that justification of the social responsibilities of computer experts is connected to more general notions of duty and responsibility, computer ethicists have drawn on a variety of traditional philosophical concepts and theories, but especially social contract theory.

Notice that the connection between being a computer expert and having a duty to deploy that expertise for the good of humanity cannot be explained simply as a causal relationship. For one thing, one can ask "why?" Why does the role of computer expert carry with it social responsibilities? For another, individuals acting in occupational roles are typically not acting simply as individual autonomous moral agents; they act as employees of companies or agencies, and may not be involved in the decisions that most critically determine project outcomes. Hence, there is a theoretical problem in explaining why and to what extent individuals acting in occupational roles are responsible for the effects of their work.

Social contract theory provides an account of the connection between occupational roles and social responsibilities. A social contract exists between members of an occupational group and the communities or societies of which they are a part. Society (states, provinces, communities) allows occupational groups to form professional organizations, to make use of educational institutions to train their members, to control admission,

and so on, but all of this is granted in exchange for a commitment to organize and control the occupational group in ways that benefit society. In other words, a profession and its members acquire certain privileges in exchange for accepting certain social responsibilities.

The substantive content of those responsibilities has also been a topic of focus for computer ethicists. Computer professional groups have developed and promulgated codes of professional and ethical conduct that delineate in broad terms what is and is not required of computer experts. See, for example, the ACM Code of Ethics and Professional Conduct or the Code of Conduct of the British Computer Society. Since these codes are very general, there has been a good deal of discussion as to their appropriate role and function. Should they be considered comparable to law? Should there be enforcement mechanisms and sanctions for those who violate the code? Or should codes of conduct aim at inspiration? If so, then they should merely consist of a statement of ideals and need not be followed "to the letter" but only in spirit.

At least one computer ethicist has gone so far as to argue that the central task of the field of computer ethics is to work out issues of professional ethics for computer professionals. Gotterbarn (1995: 21) writes that the "only way to make sense of 'Computer Ethics' is to narrow its focus to those actions that are within the control of the individual *moral* computer professional."

While Gotterbarn's position is provocative, it is not at all clear that it is right. For one thing, many of the core issues in computer ethics are social value and policy issues, such as privacy and property rights. These are issues for all citizens, not just computer professionals. Moreover, many of the core issues faced by computer professionals are not unique to computing; they are similar to issues facing other occupational groups: What do we owe our clients? Our employers? When are we justified in blowing the whistle? How can we best protect the public from risk? Furthermore, since many computer professionals work in private industry, many of the issues they face are general issues of business ethics. They have to do with buying and selling, advertising, proprietary data, competitive practices, and so on. Thus, it would be a mistake to think that all of the ethical issues surrounding computer and information technology are simply ethical issues for computer professionals. Computer experts face many complex and distinctive issues, but these are only a subset of the ethical issues surrounding computer and information technology.

3.2 Privacy

In an "information society" privacy is a major concern in that much (though by no means all) of the information gathered and processed is information about individuals. Computer technology makes possible a previously unimaginable magnitude of data collection, storage, retention, and exchange. Indeed, computer technology has made information collection a built-in feature of many activities, for example, using a credit card, making a phone call, browsing the web. Such information is often referred to as transaction-generated information or TGI.

Computer ethicists often draw on prior philosophical and legal analysis of privacy and focus on two fundamental questions: What is privacy? Why is it of value? These questions have been contentious and privacy often appears to be an elusive concept. Some argue that privacy can be reduced to other concepts such as property or liberty; some argue that privacy is something in its own right and that it is intrinsically valuable; yet others argue that while not intrinsically valuable, privacy is instrumental to other things that we value deeply – friendship, intimacy, and democracy.

Computer ethicists have taken up privacy issues in parallel with more popular public concerns about the social effects of so much personal information being gathered and exchanged. The fear is that an "information society" can easily become a "surveillance society." Here computer ethicists have drawn on the work of Bentham and Foucault suggesting that all the data being gathered about individuals may create a world in which we effectively live our daily lives in a panopticon (Reiman 1995). "Panopticon" is the shape of a structure that Jeremy Bentham designed for prisons. In a panopticon, prison cells are arranged in a circle with the inside wall of

each cell made of glass so that a guard, sitting in a guard tower situated in the center of the circle, can see everything that happens in each and every cell. The effect is not two-way; that is, the prisoners cannot see the guard in the tower. In fact, a prison guard need not be in the guard tower for the panopticon to have its effect; it is enough that prisoners believe they are being watched. When individuals believe they are being watched, they adjust their behavior accordingly; they take into account how the watcher will perceive their behavior. This influences individual behavior and how individuals see themselves.

While computerized information-gathering does not physically create the structure of a panopticon, it does something similar insofar as it makes a good deal of individual behavior available for observation. Thus, data collection activities of an information society could have the panopticon effect. Individuals would know that most of what they do can be observed and this could influence how they behave. When human behavior is monitored, recorded, and tracked, individuals could become intent on conforming to norms for fear of negative consequences. If this were to happen to a significant extent, it might incapacitate individuals in acting freely and thinking critically – capacities necessary to realize democracy. In this respect, the privacy issues around computer technology go to the heart of freedom and democracy.

It might be argued that the panoptic effect will not occur in information societies because data collection is invisible so that individuals are unaware they are being watched. This is a possibility, but it is also possible that as individuals become more and more accustomed to information societies, they will become more aware of the extent to which they are being watched. They may come to see how information gathered in various places is put together and used to make decisions that affect their interactions with government agencies, credit bureaus, insurance companies, educational institutions, employers, etc.

Concerns about privacy have been taken up in the policy arena, with a variety of legislation controlling and limiting the collection and use of personal data. An important focus here has been comparative analyses of policies in different countries – for they vary a good deal. The American approach has been piecemeal, with separate legislation for different kinds of records (i.e., medical records, employment histories, credit records), whereas several European countries have comprehensive policies that specify what kind of information can be collected under what conditions in *all* domains. Currently the policy debates are pressured by the intensification of global business. Information-gathering organizations promise data subjects that they will only use information in certain ways; yet, in a global economy, data collected in one country – with a certain kind of data protection – can flow to another country where there is no or different protection. An information-gathering organization might promise to treat information in a certain way, and then send the information abroad where it is treated in a completely different way, thus breaking the promise made to the data subject. To assure that this does not happen, a good deal of attention is currently being focused on working out international arrangements and agreements for the flow of data across national boundaries.

3.3 Cybercrime and abuse

While the threats to privacy described above arise from *uses* of computer and information technology, other threats arise from *abuses*. As individuals and companies do more and more electronically, their privacy and property rights become ever more important, and these rights are sometimes threatened by individuals who defy the law or test its limits. Such individuals may seek personal gain or may just enjoy the challenge of figuring out how to *crack* security mechanisms. They are often called *hackers* or *crackers*. The term *hacker* used to refer to individuals who simply loved the challenge of working on programs and figuring out how to do complex things with computers, but did not necessarily break the law. *Crackers* were those who broke the law. However, the terms are now used somewhat interchangeably to refer to those who engage in criminal activity.

The culture of hackers and crackers has been of interest not only because of the threat posed by their activities, but also because the culture of hackers and crackers represents an alternative

vision of how computer technology might be developed and used, one that has intrigued philosophers. (See Chapter 7 on INTERNET CULTURE.) Hackers and crackers often defend their behavior by arguing for a much more open system of computing with a freer flow of information, creating an environment in which individuals can readily share tools and ideas. In particular, the culture suggests that a policy of no ownership of software might lead to better computing. This issue goes to the heart of philosophical theories of property, raising traditional debates about the foundations of property, especially intellectual property.

Some draw on Locke's labor theory of property and argue that software developers have a natural right to control the use of their software. Others, such as me, argue that while there are good utilitarian reasons for granting ownership in software, natural rights arguments do not justify private ownership of software (Johnson 2001). There is nothing inherently unfair about living in a world in which one does not own and cannot control the use of software one has created.

Nevertheless, currently, in many industrialized countries there are laws against copying and distributing proprietary software, and computer ethicists have addressed issues around violations of these laws. Conceptually, some have wondered whether there is a difference between familiar crimes such as theft or harassment and parallel crimes done using computers. Is there any morally significant difference between stealing (copying and selling copies of) a software program and stealing a car? Is harassment via the internet morally any different than face-to-face harassment? The question arises because actions and interactions on the internet have some distinguishing features. On the internet, individuals can act under the shroud of a certain kind of anonymity. They can disguise themselves through the mediation of computers. This together with the reproducibility of information in computer systems makes for a distinctive environment for criminal behavior. One obvious difference in cybertheft is that the thief does not deprive the owner of the use of the property. The owner still has access to the software, though of course the market value of the software is diminished when there is rampant copying.

Computer ethicists have taken up the task of trying to understand and conceptualize cybercrimes as well as determining how to think about their severity and appropriate punishment. Criminal behavior is nothing new, but in an information society new types of crimes are made possible, and the actions necessary to catch criminals and prevent crimes are different.

3.4 Internet issues

Arguably the internet is the most powerful technological development of the late twentieth century. The internet brings together many industries, but especially the computer, telecommunications, and media enterprises. It brings together and provides a forum for millions of individuals and businesses around the world. It is not surprising, then, that the internet is currently a major focus of attention for computer ethicists. The development of the internet has involved moving many basic social institutions from a paper and ink medium to the electronic medium. The question for ethicists is this: is there anything ethically distinctive about the internet? (A parallel question was asked in the last section with regard to cybercrime.)

The internet seems to have three features that make it unusual or special. First, it has an unusual scope in that it provides many-to-many communication on a global scale. Of course, television and radio as well as the telephone are global in scale, but television and radio are one-to-many forms of communication, and the telephone, which is many-to-many, is expensive and more difficult to use. With the internet, individuals and companies can have much more frequent communication with one another, in real time, at relatively low cost, with ease and with visual as well as sound components. Second, the internet facilitates a certain kind of anonymity. One can communicate extensively with individuals across the globe (with ease and minimal cost), using pseudonyms or real identities, and yet one never has to encounter the others face-to-face. This type of anonymity affects the content and nature of the communication that takes place on the internet. The third special feature of the internet is its reproducibility. When put on the internet,

text, software, music, and video can be duplicated *ad infinitum*. They can also be altered with ease. Moreover, the reproducibility of the medium means that all activity on the internet is recorded and can be traced.

These three features of the internet – global many-to-many scope, anonymity, and reproducibility – have enormous positive as well as negative potential. The global, many-to-many scope can bring people from around the globe closer together, relegating geographic distance to insignificance. This feature is especially freeing to those for whom travel is physically challenging or inordinately expensive. At the same time, these potential benefits come with drawbacks; one of the drawbacks is that this power also goes to those who would use it for heinous purposes. Individuals can – while sitting anywhere in the world, with very little effort – launch viruses and disrupt communication between others. They can misrepresent themselves and dupe others on a much larger scale than before the internet.

Similarly, anonymity has both benefits and dangers. The kind of anonymity available on the internet frees some individuals by removing barriers based on physical appearance. For example, in contexts in which race and gender may get in the way of fair treatment, the anonymity provided by the internet can eliminate bias; for example, in on-line education, race, gender, and physical appearance are removed as factors affecting student-to-student interactions as well as the teacher evaluations of students. Anonymity may also facilitate participation in beneficial activities such as discussions among rape victims or battered wives or ex-cons where individuals might be reluctant to participate unless they had anonymity.

Nevertheless, anonymity leads to serious problems of accountability and for the integrity of information. It is difficult to catch criminals who act under the shroud of anonymity. And anonymity contributes to the lack of integrity of electronic information. Perhaps the best illustration of this is information one acquires in chatrooms on the internet. It is difficult (though not impossible) to be certain of the identities of the persons with whom one is chatting. The same person may be contributing information under multiple identities; multiple individuals may be using the same identity; participants may have vested interests in the information being discussed (e.g., a participant may be an employee of the company/product being discussed). When one can't determine the real source of information or develop a history of experiences with a source, it is impossible to gauge the trustworthiness of the information.

Like global scope and anonymity, reproducibility also has benefits and dangers. Reproducibility facilitates access to information and communication; it allows words and documents to be forwarded (and downloaded) to an almost infinite number of sites. It also helps in tracing cybercriminals. At the same time, however, reproducibility threatens privacy and property rights. It adds to the problems of accountability and integrity of information arising from anonymity. For example, when I am teaching a class, students can now send their assignments to me electronically. This saves time, is convenient, saves paper, etc. At the same time, however, the reproducibility of the medium raises questions about the integrity of the assignments. How can I be sure the student wrote the paper and didn't download it from the web?

When human activities move to the internet, features of these activities change and the changes may have ethical implications. The internet has led to a wide array of such changes. The task of computer ethics is to ferret out these changes and address the policy vacuums they create.

3.5 Virtual reality

One of the most philosophically intriguing capacities of computer technology is "virtual reality systems." These are systems that graphically and aurally represent environments, environments into which individuals can project themselves and interact. Virtual environments can be designed to represent real-life situations and then used to train individuals for those environments, e.g., pilot training programs. They can also be designed to do just the opposite, that is, to create environments with features radically different from the real world, e.g., fantasy games. Ethicists have just begun to take up the issues posed by virtual reality and the issues are deep (Brey 1999). The meaning of actions in virtual reality is what is at

stake as well as the moral accountability of individual behavior in virtual systems. When one acts in virtual systems one "does" something, though it is not the action represented. For example, killing a figure in a violent fantasy game is not the equivalent of killing a real person. Nevertheless, actions in virtual systems can have real-world consequences; for example, violence in a fantasy game may have an impact on the real player or, as another example, the pilot flying in the flight simulator may be judged unprepared for real flight. As human beings spend more and more time in virtual systems, ethicists will have to analyze what virtual actions mean and what, if any, accountability individuals bear for their virtual actions. (See Chapter 12 for more on VIRTUAL REALITY.)

4 Conclusion

This chapter has covered only a selection of the topics addressed by philosophers working in the field of computer ethics. Since computers and information technology are likely to continue to evolve and become further integrated into the human and natural world, new ethical issues are likely to arise. On the other hand, as we become more and more accustomed to acting with and through computer technology, the difference between "ethics" and "computer ethics" may well disappear.

Websites and Other Resources

Ethics and Information Technology, an international quarterly journal by Kluwer Academic Publishers (and the only journal devoted specifically to moral philosophy and information and communication technology; first published in 1999), contains articles on a variety of topics.

Tavani, H. 1996. "Bibliography: a computer ethics bibliography." *Computers & Society SIGCASE Reader 1996.* New York, NY: ACM Inc. This is an extremely useful resource which Tavani continues to update at: <http://www.rivier.edu/faculty/htavani/biblio.htm>.

<http://www.ethics.ubc.ca/resources/computer/>
This is the Computer and Information Ethics Resources portion of the website of the Centre for Applied Ethics of the University of British Columbia. The site includes Starting Points in Computer Ethics/Info-Tech Ethics, a set of papers to read, and a bookstore showing books in computer and information ethics which are linked to Amazon.com.

<http://www.wolfson.ox.ac.uk/~floridi/>
This website is the work of Luciano Floridi. It contains his paper entitled "Information ethics: on the philosophical foundation of computer ethics," and includes a list of resources as well as links to other projects and papers by Floridi.

<http://www.ccsr.cse.dmu.ac.uk/contents/>
This is the site for the Centre for Computing and Social Responsibility (CCSR), in the UK, provides access to a variety of useful materials including a list of conferences, a discussion forum and links to other sites.

<http://onlineethics.org>
This website is devoted broadly to engineering and computer ethics, and contains bibliographic materials and case studies, as well as links to other sites.

References

Adams, A. and Ofori-Amanfo, J. 2000. "Does gender matter in computer ethics?" *Ethics and Information Technology* 2(1): 37–47.

Anderson, T. C. 2000. "The body and communities in cyberspace: a Marcellian analysis." *Ethics and Information Technology* 2(3): 153–8.

Asimov, I. 1970. *I, Robot.* Greenwich, CT: Fawcett Publications.

Brey, P. 1999. "The ethics of representation and action in virtual reality." *Ethics and Information Technology* 1(1): 5–14.

Brey, P. 2000. "Disclosive computer ethics." *Computers & Society*, Dec.: 10–16.

Castells, M. 1996. *The Rise of the Network Society.* Malden, MA: Blackwell Publishers.

——. 1997. *The Power of Identity.* Malden, MA: Blackwell Publishers.

——. 1998. *The End of Millennium.* Malden, MA: Blackwell Publishers.

Cohen, R. A. 2000. "Ethics and cybernetics: Levinasian reflections." *Ethics and Information Technology* 2(1): 27–35.

Dibbell, J. 1993. "A rape in cyberspace: how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society." *The Village Voice*, Dec. 23: 36–42.

Dreyfus, H. L. 1999. "Anonymity versus commitment: the dangers of education on the internet." *Ethics and Information Technology* 1(1): 15–21.

Ess, C., ed. 1996. *Philosophical Perspectives on Computer-Mediated Communication*. Albany: State University of New York Press.

Floridi, L. 1999. "Information ethics: on the philosophical foundation of computer ethics." *Ethics and Information Technology* 1(1): 37–56.

Gotterbarn, D. 1995. "Computer ethics: responsibility regained." In D. G. Johnson and H. Nissenbaum, eds., *Computers, Ethics and Social Values*. Englewood Cliffs, NJ: Prentice Hall, pp. 18–24.

Grodzinsky, F. S. 1999. "The practitioner from within: revisiting the virtues." *Computers & Society* 29(1): 9–15.

Introna, L. D. 2001. "Proximity and simulacra: ethics in an electronically mediated world." *Philosophy in the Contemporary World* (forthcoming).

—— and Nissenbaum, H. 2000. "Shaping the web: why the politics of search engines matters." *The Information Society* 16(3): 169–85.

Johnson, D. G. 2001. *Computer Ethics*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Moor, J. 1985. "What is computer ethics?" *Metaphilosophy* 16(4): 266–75.

Prosser, B. T. and Ward, A. 2000. "Kierkegaard and the internet: existential reflections on education and community." *Ethics and Information Technology* 2(3): 167–80.

Reiman, J. H. 1995. "Driving to the panopticon: a philosophical exploration of the risks to privacy posed by the highway technology of the future." *Computer and High Technology Law Journal* 11: 27–44.

van den Hoven, J. 1999. "Privacy and the varieties of informational wrongdoing." *Australian Journal of Professional and Applied Ethics* 1(1): 30–43.

Chapter 6

Computer-mediated Communication and Human–Computer Interaction

Charles Ess

Introduction: CMC and Philosophy

From Anaximander through Kant, philosophers have recognized that knowing a thing involves knowledge of its limits, i.e., the boundaries or edges that define (delimit) both what a thing is and what it is not. Information and Computing Technologies (ICT) give philosophers powerful new venues for examining previously held beliefs concerning what *delimits* human beings, for instance, artificial *vis-à-vis* natural intelligence. As we will see, ICT further allow us to test long-debated claims regarding human nature and thus politics, that is, questions like whether we are capable of democratic governance or we require authoritarian control. Computer-Mediated Communication (CMC) and Human–Computer Interaction (HCI) provide philosophers with new laboratories in which claims that previously rested primarily on the force of the best arguments can now be reevaluated empirically, in light of the attempts made to implement these assumptions in the *praxis* of human–machine interaction, in the potentially democratizing effects of CMC, and so forth (on this new methodological turn see also Chapter 26, Philosophy of Information Technology).

To see how this is so, this chapter begins with some elementary definitions. The second section provides an analysis of some of the key philosophical issues that are illuminated through various disciplinary approaches that incorporate CMC technologies. This discussion is organized in terms of the fundamental elements of *worldview*, i.e., of ontology, epistemology (including semiotics, hypertext, and logic), the meaning of identity and personhood (including issues of gender and embodiment), and ethical and political values (especially those clustering about the claim that these technologies will issue in a global democracy vs. the correlative dangers of commercialization and a "computer-mediated colonization"). In the last section, some suggestions of possible research directions for, and potential contributions of, "computer-mediated philosophy" are offered, in view of a philosophical inquiry oriented towards the sorts of theories, *praxis*, and interdisciplinary dialogues described here. Perhaps most importantly, philosophers may be able to contribute to a renewed education – one taking

Socrates as its model – that is required for cultural flows facilitated by CMC.

1 Some Definitions

CMC may be defined as interactive communication between two or more intelligent agents that relies on ICT – usually personal computers and networks – as its primary medium. Examples include: e-mail, chatrooms, USENET newsgroups, MUDs and MOOs, listserves, "instant messaging" services (ICQ, AOL Instant Messager, etc.), audio- and video-teleconferencing, shared virtual reality systems, and other ways of sharing files and information via networks and the internet, including peer-to-peer file transfers (via a service such as Gnutella, <http://www.gnutella.com>), and the multimedia communication of the web (e.g., personal homepages, folder- and link-sharing via <http://www.backflip.com>, photo-file sharing on commercial servers, etc.). This definition allows for the possibility of humans communicating with intelligent but artificial agents via computers and networks and, as we will see below, thus points towards artificial intelligence and related developments as limiting issues for CMC (see Herring 2002 for a more complete description and history of the most significant examples of CMC).

HCI may be construed as a narrowly defined variant of CMC. While CMC refers to any communication between intelligent agents mediated by computers, such communication usually includes, and thus presupposes, successful interaction between human agents and the mediating technologies. Such interaction requires an interface design that, ideally, allows for "seamless" or "intuitive" communication between human and machine. The design of such interfaces, and the correlative investigations into human and machine capacities, cognitive abilities, and possible ways of interacting with the world and one another constitute the subject matter of HCI. While HCI is incipient in every computer design, early HCI literature largely assumed that machines would be used by an elite of technical experts; but as computing technologies became more ubiquitous, so the need increased for more "user-friendly" interface design, thus requiring greater attention to HCI issues (Bardini 2000, Hollan 1999, Suchman 1987).

Finally, as Carleen Maitland (2001) points out, the research area of computer-supported cooperative work (CSCW) may be included as a subarea of CMC/HCI.

2 Philosophical Perspectives: Worldview

While extensive and growing almost as explosively as the internet and the web themselves, both scholarly and popular literatures in CMC, HCI, and CSCW remain primarily within the boundaries of the disciplines of computer science, "human factors" as understood in terms of ergonomics, communication theory, cultural studies, and such social sciences as ethnography, anthropology, psychology, and, especially, in the case of CSCW, the social psychology of group work (Hakken 1999; Bell 2000). Some theorists and designers exploit the theoretical frameworks and insights of cognitive psychology, cognitive science, artificial intelligence, and so forth, thus approaching more directly philosophical domains. Finally, some examples represent an explicit dialogue between CMC and HCI on the one hand, and philosophical concerns on the other. The communication theorists Chesebro and Bertelson, for example, utilize a theory of communication originally developed by Innis, Eisenstein, McLuhan, and Ong, that sees communication as a technology that in turn centrally defines culture, in order to explicitly address philosophical concerns with epistemology, ontology, critical reasoning, etc. (Chesebro & Bertelson 1996; see Ess 1999). Taken together, these contribute significantly to the characteristically philosophical projects of uncovering and articulating basic *worldview* assumptions such as *epistemology* (including questions concerning the nature of truth, whether truth may have a universally valid status, etc.), *ontology* (including questions concerning the reality and meaning of *being human*), *ethics*, *politics* (including issues of democracy and justice), and so forth

(see also in this volume Chapters 11 and 12, Ontology and Virtual Reality).

2.1 Ontology, epistemology, personhood

In this chapter the term "ontology" is used in a broad sense, one that includes more traditional metaphysics. This category raises questions about the nature of the real, including both internal entities (such as a self, mind, and/or spirit), and external realities as well as an external world or worlds, including persons, transcendental realities (mathematical, ethical [e.g., values and rights that are not reducible to the strictly material], religious, etc.), causal and other possible relationships.

Beyond questions regarding ontology and virtual reality, questions concerning human nature and the self are among the most prominent ontological questions evoked by, and explored in, CMC and HCI. These questions are perhaps as ancient as speculation concerning the Golem and automata in the fifteenth and sixteenth centuries. In any case, directions for design of HCI were defined from the 1950s on by two distinct philosophical visions. The first (originally, the minority position represented by Douglas Engelbart) was a more humanistic – indeed, classically Enlightenment/Cartesian – vision of using computing technologies as slaves, in a symbiosis intended to *augment*, not *replace*, human intelligence. The second (originally more dominant) vision of the AI community was to build superior replacements of the human mind. This general project is commonly characterized by a Cartesian dualism, one that regards the mind as reason divorced from body and whose primary mode of knowledge is mathematical and symbolic (see, however, Floridi 1999 for a more extensive analysis of the philosophical assumptions underlying so-called strong AI, one that argues against the view that AI rests on Cartesian roots). The former emphasized the need for HCI design to accommodate the machine to the human by recognizing that the machine *differs* from the human in important ways. Its binary language and symbolic processes do not neatly match human natural language, and the human "interface" with our world includes that of an *embodied* mind, one whose interaction with the machine will thus turn on a variety of physical devices (most famously, Engelbart's mouse) and multiple senses (including a graphical user interface that exploits the visual organization of information). The AI orientation tended to minimize matching human and computer in terms of interface, partly because any human–machine symbiosis was seen as only an intermediate stage on the way to machines replacing human beings (Bardini 2000: 21). Engelbart's "coevolutionary" approach to HCI, by contrast, rests on an analogous dialogue between disciplines. He was directly influenced by linguist Benjamin Whorf and the recognition of the role of natural language in shaping *worldview* (Bardini 2000: 36). Worldview is thus the conceptual interface between HCI, linguistics, and philosophy.

Winograd and Flores (1986) more explicitly take up the philosophical dimensions of the split in HCI between AI and Engelbart. They explore the intersections between computer technology and the nature of human existence, including "the background of a tacit understanding of human nature and human work." They clarify that "in designing tools we are designing ways of being" (1986: xi). That is: tools are designed with the goal of making specific actions and processes easier and thus their design reflects a range of assumptions, including worldview assumptions regarding what is valuable, what is possible and easy for the users involved, and what are the preferred ways of facilitating these processes. As they make certain actions and processes easier, tools thus embody and embed these assumptions, while excluding others. In doing so, they thus bias their users in specific directions and, in this way, shape our possible ways of being. Following Bardini's analysis of the dominance of AI-oriented approaches in earlier HCI, Winograd and Flores interpret the worldview of much computer design as "rationalistic," "because of its emphasis on particular styles of consciously rationalized thought and action" (1986: 8). They seek to correct its "particular blindness about the nature of human thought and language" by highlighting how language and thought depend on social interaction, an analysis based on the philosophical traditions of hermeneutics and phenomenology and including Heidegger, Austin, Searle, and Habermas (1986: 9).

Winograd and Flores's project of unveiling the established but tacit background knowledge of computer designers regarding *what it means to be human* anticipates a burgeoning discussion in CMC, HCI, and CSCW literatures in the 1990s concerning specific conceptions of personhood and identity presumed by various design philosophies. A central focal point for this discussion is the notion of the cyborg, the human–machine symbiosis originally figuring in science fiction, perhaps most prominently as the Borg in *Star Trek: The Next Generation*. The Borg can represent humanity's worst fears about technology. Once the boundary between humanity and machinery is breached, the machinery will irresistibly take control, destroying our nature (specifically, the capacities for independent agency and compassion towards others) in the process. By contrast, Donna Haraway's "Cyborg Manifesto" (1990) argues that women as embodied creatures are thus trapped in a real world of patriarchal oppression, one in which women, body, and sexuality are demonized. Women (and men) can thus find genuine equality and liberation only as disembodied minds in cyberspace, as cyborgs liberated rather than dehumanized through technology.

Philosophers will recognize in Haraway's vision of technologically mediated liberation a dualism that echoes Descartes' mind–body split. For historians of religion, such dualism further recalls Gnostic beliefs. Gnostics held that the human spirit is a kind of divine spark fallen from heaven and trapped within the debased materiality of the human body. For such a spirit – as ontologically and ethically opposed to the body – salvation can come only through liberation from the body. Such Gnosticism appears to be at work in numerous visions of liberation through CMC technologies, including explicitly religious ones (O'Leary & Brasher 1996; Wertheim 1999). As Katherine Hayles (1999) has documented, this dualism emerges in the foundational assumptions of cybernetics and a conception of formalistic rationality in AI, one that issues most famously in Hans Moravec's hope that humans will soon be able to download their consciousness into robotic bodies that will live forever (1988). This dualism, moreover, can be seen at work in the relatively early celebration of hypertext and CMC as marking out a cultural shift as revolutionary as the printing press, if not the invention of fire (e.g., Lyotard 1984, Bolter 1984, 1991, Landow 1992, 1994). That is, to emphasize the radical difference between print culture and what Ong has called the "secondary orality" of electronic media and culture (1988) requires us to establish a dualistic opposition between these two cultural stages, one fostered by especially post-modernist emphases on such a radical dichotomy between modernity and postmodernity. This emphasis on the radical/revolutionary difference between past and future is, precisely, consistent with Haraway's early "cyber-gnosticism," the equally dualistic presumption that the mind/persona in cyberspace is radically divorced from the body sitting back at the keyboard. Such cyber-gnosticism takes political expression in the libertarian hopes for a complete liberation from the chains of modernity and the constraints of what John Perry Barlow so contemptuously called "meatspace" (1996).

The difficulties of dualism and Gnosticism, however, are well known, ranging from the mind–body problem (in Descartes's terms, how does mind as a thinking, non-extended substance communicate with and affect the body as a non-thinking, extended substance?) to what Nietzsche identified as "the metaphysics of the hangman," i.e., the objection that especially Christian dualisms result in a denigration of body, sexuality, women, and "this life" in general (1954: 500). In light of these classical difficulties, the more recent turn from such dualisms in the literatures of CMC and HCI is not surprising. To begin with, alternatives to the Cartesian/AI conceptions of knowledge began to emerge within the literatures of cybernetics and HCI, e.g. in Bateson's notion of distributed cognition (1972, 1979) and Engelbart's emphasis on kinesthetic knowledge (Bardini 2000: 228f.; Hayles 1999: 76–90; cf. Suchman 1987). A more recent example of this turn is Hayles' version of the "posthuman," as characterized by an explicit epistemological agenda: "reflexive epistemology replaces objectivism . . . embodiment replaces a body seen as a support system for the mind; and a dynamic partnership between humans and intelligent machines replaces the liberal humanist subject's manifest destiny to dominate and control nature" (1999:

288). That is, Hayles foregrounds here the shift from an objectivist epistemology, based on a dualistic separation of subject–object (and thus between subjective vs. objective modes of knowledge, so as to then insist that only "objective" modes of knowledge are of value), to an epistemology which (echoing Kant) emphasizes the inevitable interaction between subject and object in shaping our knowledge of the world. Knowledge is not an "either/or" between subjective and objective, it is *both* subjective *and* objective. In the same way, Hayles further focuses precisely on the meanings of *embodiment* in what many now see as a post-Cartesian understanding of mind-*and*-body in cyberspace (Bolter 2001; Brown & Duguid 2000; Dertouzos 2001). These shifts, finally, undercut the Cartesian project of using technology to "master and possess nature" (Descartes 1637: 35). Such a project makes sense for a Cartesian mind radically divorced from its own body and nature, indeed, a mind for whom nature is seen as inferior and dependent (1637: 19). But as the environmental crises of our own day make abundantly clear, as embodied beings (not just "brains on a stick") we *are* intimately interwoven with a richly complex natural order, *contra* the Cartesian dualisms underlying what Hayles calls the liberal project of mastery and domination of nature. More broadly, especially as the demographics of the Net change, and women are now the majority of users, it seems likely that the literature on gender, cyborgs, and personhood will continue to offer new philosophical insights and directions.

There emerges here then a series of debates between postmodern/dualistic emphases on radical *difference* between mind and body, humanity and nature, electronic and print cultures, etc., and more recent reconsiderations that stress *connection* between these dyadic elements. These debates are further at work in philosophical considerations of *space* and *place*. On the one hand, the very term "cyberspace" indicates that our ordinary conceptions cannot fully apply to the new sorts of individual and social spaces enabled by these technologies. Similarly, Mike Sandbothe (1999), partly relying on Rorty and Derrida, has argued that the internet and the web collapse "natural" senses of time into the virtually instantaneous, thus making the experience of time one shaped by users. Especially given a postmodernist or social-constructivist epistemology that minimizes the role of any external givens as constraining our knowledge, time and space may become our own creations, the result of aesthetic choices and our narrative and cooperative imagination.

On the other hand, the renewed stress on the ontological/epistemological connections between mind and world and the corresponding ethical and political responsibilities entailed by such connections further parallel observations that, *contra* the ostensibly transnational character of the web and the Net, social and national boundaries are in fact observed in cyberspace (Halavais 2000), with potentially imperialistic consequences (Barwell & Bowles 2000). As we will see in the discussion of politics, the strength of the *connections* between physical spaces and cyberspace is further apparent if we examine the role of diverse cultures in resisting a "computer-mediated colonialism," i.e., the imposition of Western values and communication preferences across the globe as these values and preferences are embedded in the current technologies of CMC and CSCW. Recent work documents the many ways in which especially Asian cultures – whose cultural values and communicative preferences perhaps most clearly diverge from those embedded in Western CMC and CSCW technologies – are able to reshape CMC and CSCW technologies in order better to preserve and enhance distinctive cultural values and identity.

2.2 Epistemology: semiotics, hypertext, and logic

The notion of "communication" in CMC combines philosophical and communication theoretical views. For example, Shank and Cunningham (1996) argue that CMC requires moving from a Cartesian view, according to which autonomous minds transfer information across a transparent medium, to a theoretical approach reflecting both communication theories that stress intersubjectivity (as instantiated in dialogues and multilogues, in contrast with a monologue) and C. S. Peirce's semiotics, which emphasizes the emergence of meaning out of an interconnected triad of objects, signs, and persons (or, more generally what Peirce

calls "interpretants"). Peirce remains an important point of dialogue between philosophers and CMC theorists (Mayer 1998, Groleau & Cooren 1999).

Chapter 19 takes up *hypertext* in greater detail. Here we can note that David Kolb (1996) has explored how hypertextual CMC technologies may preserve and expand the discursive moves of argument and criticism, in a domain that is both hypertextual and, using Ong's terms, "oral" (i.e., ostensibly marked by greater equality and participation than in the more hierarchical societies of literate and print cultures) *vis-à-vis* the ostensive linearity of print. Against the postmodern emphasis on hypertext as radically overturning ostensibly modern and/or solely literate modes of reasoning and knowledge, Kolb argues that hypertext can facilitate especially the dialectical patterns of Hegelian and Nietzschean argument. But *contra* postmodern celebrations of hypertext as exploding all print-based constraints, Kolb emphasizes the reality of humans as *finite* creatures. In the face of the "information flood" of an exponentially expanding web of argumentative hypertexts, Kolb (rightly) predicts that the finitude of human knowers will require new centers of access to exploding information flows, thus engendering new forms of hypertextual discourse.

Herbert Hrachovec (2001) has explored CMC as a potential "space of Reason," one whose hypertextual dimensions either (a) reinstantiate traditional print-based modes of knowledge representation and argument (the Talmud, indices, cross-references, use of images in medieval manuscripts, etc.) and/or (b) fundamentally challenge and surpass traditional forms of knowledge, argument, and reason (for similar sorts of discussion concerning how computer technologies may reshape received notions of logic, see Scaltsas 1998, Barwise & Etchemendy 1998).

Some famous (but controversial) studies have documented negative social consequences correlating with increased participation in cyberspace. Even such prominent proponents as Jay David Bolter (2001) acknowledge that electronic environments favor the personal and playful rather than abstract reasoning. These observations raise additional questions as to how CMC technologies may be shaping consciousness in ways potentially antiphilosophical, or at least "differently" philosophical. For example, if we live increasingly in a style of multitasking and "partial attention" (Friedman 2001), how well will complex philosophical arguments requiring sustained intellectual attention remain accessible to novice philosophers? Similarly, traditional philosophical conceptions of the self include a singular agent, as a moral agent responsible for its acts over time or as an epistemological agent, such as Kant's transcendental unity of apperception, whose unitary nature is inferred from the coherence of an experiential stream of sense-data that otherwise tends to scatter centrifugally. Of course, postmodernism counters with notions of multiple, decentered, fragmented selves. Postmodernist theories dominated early CMC literature, celebrating the hypertextual web of cyberspace precisely as it appeared to instantiate such conceptions of self. Should our immersion into cyberspace issue exactly in such decentered selves, however, the philosophical debates between modernists and postmodernists concerning the self may become irrelevant. Selves that are *de facto* decentered and fragmented would be incapable of the sustained attention required for complex philosophical arguments – as well as incapable of acting as singular epistemological and moral agents. Such selves would *not* demonstrate the cogency of the postmodern concept as resulting from rigorous philosophical debate between moderns and postmoderns. Rather, such selves would represent only a technologically aided self-fulfilling prophecy, i.e., the result of adopting such technologies in the first place because we uncritically and without further argument presume the truth of the postmodern notions of self as justification for immersing ourselves in the technologies that produce such selves. As this last phrase tries to make clear, such self-fulfilling prophecies are, in logical terms, viciously circular arguments, for their conclusions are already asserted in their premises. At stake in the debate, however, is nothing less than our most fundamental conceptions of what it means to be a human and/or a person. Both these conceptions and the consequences of uncritically accepting a given (e.g., postmodernist) conception over another are too important to have them decided for us on the basis of circular argument

and self-fulfilling prophecy, rather than through more logically sound philosophical debate.

2.3 Ethics and politics: democratization vs. the panopticon and modernism vs. postmodernism

Perhaps the single most important claim made in the effort to legitimate – if not simply sell – CMC technologies is that they will democratize, in the sense of flatten, both local (including corporate) and global hierarchies, bringing about greater freedom and equality. These claims obviously appeal to Western – specifically, both modern liberal and postmodernist – values, but require philosophical scrutiny. To begin with, much CMC and popular literature assumes that "democracy" means especially a libertarian form of democracy, in contrast with communitarian and pluralist forms (Ess 1996: 198–202; Hamelink 2000: 165–85). Much of the theoretically informed debate turns on especially Habermasian conceptions of democracy, the public sphere, and a notion of communicative reason which, coupled with the rules of discourse, may achieve, in an ideal speech situation, the freedom, equality, and critical rationality required for democracy (Ess 1996: 203–12; Hamelink 2000: 165–85). Seen as simply a final expression of modern Enlightenment, however, Habermas is criticized by feminists and postmodernists for attempting to save a notion of reason that, at best, may be simply a male form of "rationality" and, at worst, contrary to its intentions to achieve freedom and equality, threatens instead to become its own form of totalitarian power (e.g. Poster 1997: 206–10). Habermas responds to these critiques by incorporating especially feminist notions of solidarity and perspective-taking, and by criticizing postmodernism in turn as ethically relativistic and thus unable to sustain its own preferences for democratic polity over other forms (Ess 1996: 212–16; Hamelink 2000: 55–76). More recent debate between Habermas and Niklas Luhmann further sharpens the theoretical limitations of the former's conception of democracy and the public sphere. Habermas's conception of "partial publics" (*Teilöffentlichkeiten*) survives here as something of a theoretical compromise between a full-fledged public sphere on the internet and its complete absence in a postmodernist emphasis on fragmentation and decentering (Becker & Wehner 2001; cf. Jones's conceptions of "micropolis" and "compunity," 2001: 56–7; Stevenson 2000).

Examining how CMC technologies are implemented in *praxis* further illuminates this debate, where the emphasis on testing theory by attempting to realize it precisely within the particulars of everyday life is itself a Habermasian – indeed, Aristotelian – requirement. Specific instances of decision-making facilitated by CMC technologies appear to approximate the ideal speech situation and realize at least a partial public sphere (Ess 1996: 218–20; Becker & Wehner 2001; Sy 2001). At the same time, however, counterexamples abound, including cases of CMC technologies serving authoritarian ends and preserving cultural hierarchies of power, status, privilege, etc. (Yoon 1996: 2001). There are also middle grounds, with examples of CMC technologies leading to partial fulfillment of hopes for democracy and equality in cultural contexts previously marked by more centralized and hierarchical forms of government (Dahan 1999, Hongladarom 2000, 2001, Wheeler 2001). These diverse results suggest that realizing the democratic potentials of CMC will require conscious attention to the social context of use, including education, a point we shall return to below.

2.4 Globalization, commercialization, and commodification vs. individual, local identity

Economic and infrastructure realities dramatically call into question the assumption that CMC represents a democratizing technology insofar as it is interactive and can place a printing press in the hands of anyone who can afford a computer and internet access. Currently, less than 7 percent of the world's population enjoys such access (see <http://www.nua.ie/surveys/how_many_online/>). Commercialization and commodification work against any such democratization effect (Poster 1997, Stratton 1997, McChesney 2000, Willis 2000, Yoon 2001; see Plant 2000 for a discussion of Irigaray's notion of the

commodification of women in the "specular economy"). In particular, Sy (2001) describes the "commodification of the lifeworld," drawing on Habermas to understand how CMC technologies in the Philippines threaten to override local cultural values and communication preferences. This is a process now well-documented for numerous cultures. In the context of India, for example, Keniston (2001) analyzes commodification and other forces contributing to an emerging, culturall homogenous "McWorld," a threat to local and regional identity that understandably evokes sometimes violent but fragmenting efforts of preservation (Sardar 2000). Hamelink (1999) refers to this process as the "Disneyfication scenario" (cf. Bukatman 2000).

Nevertheless, recent research shows how local or "thick" cultures both resist a computer-mediated colonization of the lifeworld and reshape extant CMC and CSCW technologies to better preserve and enhance distinctive communicative preferences and cultural values. In the literature of CSCW, for example, Lorna Heaton (2001) documents how Japanese CSCW researchers developed their own CSCW technologies to capture the many elements of nonverbal communication crucial in Japanese culture (gesture, gaze, etc.). Similarly, in Thailand (Hongladarom 2000, 2001) and the Philippines (Sy 2001) it appears that any emerging global culture remains "thin" in Walzer's sense, i.e., it provides no sense of historical/spatial location nor any of the "thick" moral commitments and resources that distinguish the practices and preferences of one culture from the next (cf. Hamelink 1999). The dangers and problems of globalization, especially as fostered by the rapid diffusion of CMC technologies – including the presumption of a consumerist, resource-intensive, and thus non-sustainable lifestyle – are not to be dismissed. However, *contra* the claims of technological determinism, these and similar reports suggest that CMC technologies will not inevitably overrun diverse cultural values and preferences. Rather, especially when implemented in ways that attend to the social context of use, including education, these technologies may be appropriated by diverse cultures in ways that make both global (but "thin") communication and culture possible without compromising local/"thick" cultural values and preferences (e.g., Harris et al. 2001).

3 Interdisciplinary Dialogue and Future Directions in Philosophy

Philosophers have much to gain from the theory and praxis of the many disciplines clustered about CMC technologies. Despite fledgling (Ess 2001) and more considered work (Borgman 1984, 1999, Graham 1999), philosophers yet have much to contribute to an interdisciplinary dialogue with theorists and practitioners in CMC. The following is only a brief overview of three key areas of research.

3.1 Critical reflection and history of ideas

To begin with, philosophers can extend – and, when necessary, amplify and challenge – the developing histories and conceptual frameworks of CMC, especially as these intersect questions of epistemology, ethics, and ontology. Researchers in communication theory, cultural studies, HCI, etc., are not as fully versed in the history of ideas and the often complex arguments more familiar to philosophers. These limitations can result in lacunae, oversimplifications, and errors of fact and logic that philosophers can amend, thereby adding greater accuracy and conceptual strength to the discussion and development of CMC. Specifically, beyond issues of epistemology, embodiment, and what it means to be a person, philosophers may also contribute to the related theoretical-metatheoretical issue of what we mean by *culture* (see Ess 2001: 20–2).

3.2 Uncovering worldview

CMC technologies force us to articulate and, perhaps, alter and transcend the most basic elements of our worldview, including our presumptions of identity, ontology, and epistemology (Sandbothe 1999). At the same time, the abandoning of Cartesian dualism in early Haraway

and Barlow involves a renewed interest in phenomenological and hermeneutical approaches that emphasize connectedness between body and mind and between the individual and a larger community as shaped by history, tradition, culture, etc. Thus, Paul Ricoeur is enjoying a new currency (Richards 1998; Bolter 2001), as are Husserl and Nozick (McBeath & Webb 2000). In this light, Winograd and Flores, in their appeal to the hermeneutical/phenomenological philosophies of Gadamer and Heidegger, were considerably ahead of their time.

3.3 Contributing to global dialogue

Sandbothe (1999) takes up Rorty's hope that the new media may lead to a transcultural communication, one that will help us become more empathic, understanding, and receptive towards others. Sandbothe argues that as internet communication forces users to articulate our most basic assumptions about identity, time, and space, it thereby also helps us recognize the *contingent* (i.e. *non*-universal) character of these most basic presumptions. Such communication thereby issues in a kind of epistemological humility. This should short-circuit ethnocentrisms that otherwise root both tacit and overt forms of cultural imperialism, thereby contributing to the genuine dialogue across and between cultures required for the much-prophesied global village of online communities that extend beyond specific cultural boundaries (see also Ess 2001).

3.4 Education for an intercultural global village?

By engaging in an interdisciplinary theory and *praxis* of CMC, philosophers may contribute to a specific sort of *education* for the citizens of an intercultural electronic village that is required to avoid the cultural homogenization of McWorld and the radical fragmentation of Jihad.

While Plato (at least in a straw-man form) is routinely targeted especially by postmodernist critics for an alleged dualism that then grounds subsequent dualisms in Western thought, one can argue that his allegory of the cave in the *Republic* remains a vital metaphor for both philosophy and education as processes of making tacit assumptions explicit and thereby enabling a critical examination of worldview. Philosophical education moves us from the ethnocentrism of the cave to more encompassing and finally dialogical conceptions of human beings (Ess 2001). Cees Hamelink (2000: 182ff.), in his many recommendations for how to democratize technology choices, calls for an explicitly "Socratic education," one that stresses critical thinking about the risks of deploying information and communication technologies. Hamelink appeals for an education that will "prepare people for the 'culture of dialogue' that the democratic process requires," a (partially Habermasian) dialogue that will be based on citizens' "capacity to reason through their own positions and justify their preferences" as they jointly "deliberate and reflect on the choices that optimally serve the common interest" (184). Drawing on John Dewey and Martha Nussbaum, Hamelink sees such education as vital to sustaining a democratic society as now centrally engaged with the technologies of CMC. One could add that such education is simultaneously vital to any hopes for intercultural dialogue and democracy on a global scale. In addition to historical and conceptual metaphors of the postmodern and posthuman, philosophical education in intercultural values may contribute to a new Renaissance of cultural flows facilitated by CMC technologies in dramatic new ways.

References, Resources

Bardini, T. 2000. *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing*. Stanford: Stanford University Press. [A highly readable and important account of the history and philosophies behind some of the most important debates concerning the development of HCI, including those elements such as the mouse and the Graphic User Interface originally developed by Engelbart and now taken for granted by most personal computer users. Undergraduates, graduates.]

Barlow, J. 1996. "A declaration of the independence of cyberspace," <http://www.eff.org/pub/

Censorship/Internet_censorship_bills/barlow_0296.declaration>. [A frequently-cited expression of the libertarian insistence on individual freedom from all constraints – first of all, with regard to free speech – that dominated US internet culture and thus the mid-1990s internet and related CMC writing. Undergraduates, graduates.]

Barwell, G. and Bowles, K. 2000. "Border crossings: the internet and the dislocation of citizenship." In D. Bell and B. Kennedy, eds., *The Cybercultures Reader*. London and New York: Routledge, pp. 703–11. [Succinctly raises the question of whose cultures will be lost if all cultural difference is erased (the underside of the promise of the Net to eliminate the boundaries between here and there, self and Other, etc.). Undergraduates, graduates.]

Barwise, J. and Etchemendy, J. 1998. "Computers, visualization, and the nature of reasoning." In T. Bynum and J. Moor, eds., *The Digital Phoenix: How Computers are Changing Philosophy*. Oxford: Blackwell, pp. 93–116. [A foundational account by two pioneers in the use of computers to assist teaching formal logic, in part through visualization of formal relationships. Undergraduates, graduates.]

Bateson, G. 1972. *Steps to an Ecology of Mind*. New York: Ballantine Books. *Also*: Bateson, G. 1979. *Mind and Nature: A Necessary Unity*. New York: Bantam Books. [Two foundational volumes in the development of connectionist epistemologies. Undergraduates, graduates.]

Becker, B. and Wehner, J. 2001. "Electronic networks and civil society: reflections on structural changes in the public sphere." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 65–85. [Provides an excellent overview of the extensive research and scholarship pertinent to Habermas's conception of the public sphere and how far it may be instantiated in CMC environments. Advanced undergraduates, graduate students.]

Bell, D. 2000. "Approaching cyberculture: introduction." In D. Bell and B. Kennedy, eds., *The Cybercultures Reader*. London and New York: Routledge, pp. 25–8. [An introduction to five essays on cyberculture, identifying representative approaches and viewpoints. Undergraduates, graduates.]

Bolter, J. D. 1984. *Turing's Man: Western Culture in the Computer Age*. Chapel Hill: University of North Carolina Press. [One of the first and still most influential cross-disciplinary explorations of CMC, written by a classics scholar with an advanced degree in computer science. Undergraduates, graduates.]

——. 1991. *Writing Space: The Computer, Hypertext, and the History of Writing*. Hillsdale, NJ: Lawrence Erlbaum. [A central document in the prevailing arguments that the structures of hypertext conspicuously cohere with the then-emerging postmodernist themes (decentering, fragmentation, etc.). Undergraduates, graduates.]

——. 2001. "Identity." In T. Swiss, ed., *Unspun*. New York: New York University Press, pp. 17–29. Available online: <http://www.nyupress.nyu.edu/unspun/samplechap.html>. [A highly readable update of Bolter's views, including a clear attack on a Cartesian sense of self as affiliated with the culture of print vs. a postmodern "fluid cognitive psychology" affiliated with identity on the web – issuing in the insistence that CMC cannot undermine our identities as embodied. Undergraduates, graduates.]

Borgmann, A. 1984. *Technology and the Character of Contemporary Life*. Chicago: University of Chicago Press. [One of the most significant volumes in philosophy of technology; Borgmann offers his notion of "focal practices" to help offset what he argues are the humanly debilitating consequences of increasing reliance on technology. Advanced undergraduates, graduates.]

——. 1999. *Holding onto Reality: The Nature of Information at the Turn of the Millennium*. Chicago: University of Chicago Press. [One of the most significant contributions to philosophically coming to grips with "the Information Age." Borgmann seeks to develop a theory of information – first of all, by developing his own theory of signs and kinds of information – in order to then establish an ethics of information intended to help us recover "the good life" (in the deepest philosophical sense). Advanced undergraduates, graduates.]

Brown, J. S. and Duguid, P. 2000. *The Social Life of Information*. Stanford: Stanford Uuniversity Press. [A thoughtful debunking of the "techno-enthusiasm" that prevails in especially US-centered discourse concerning "information" and the Information Age – notable because its authors enjoy exceptional credentials: in particular, Brown is chief scientist at Xerox and director of its Palo Alto Research Center. Undergraduates, graduates.]

Bukatman, S. 2000. "Terminal penetration." In D. Bell and B. Kennedy, eds., *The Cybercultures Reader*. London and New York: Routledge, pp. 149–74. [Uses contemporary Continental philosophy – including Merleau-Ponty – to develop an account of the "phenomenal body" that emerges in cyberspace and virtual reality, *vis-à-vis* significant science-fiction and cyberpunk texts and movies, as well as Disney's Epcot Center. Advanced undergraduates, graduates.]

Chesebro, J. W. and Bertelsen, D. A. 1996. *Analyzing Media: Communication Technologies as Symbolic and Cognitive Systems*. New York: Guilford Press. [An important contribution towards an interdisciplinary critical theory of media, including CMC. Advanced undergraduates, graduates.]

Communication Institute for Online Research, <http://www.cios.org>. [CIOS offers a number of comprehensive databases in communication theory and research; a modest subscription fee is required for full-text access. Undergraduates, graduates.]

Dahan, M. 1999. "National security and democracy on the Internet in Israel." *Javnost – The Public* VI(4): 67–77. [Documents the role of the internet in contributing to greater openness in Israeli society. Undergraduates, graduates.]

Dertouzos, M. 2001. *The Unfinished Revolution: Human-centered Computers and What They Can Do For Us*. New York: HarperCollins. [Dertouzos (head of the MIT Laboratory for Computer Science and a thoughtful commentator) argues optimistically that a "human-centric" approach to computing, coupled with additional advances, will overcome contemporary feelings of enslavement to the machines. Undergraduates, graduates.]

Descartes, Rene. [1637] 1998. *Discourse on Method and Meditations on First Philosophy*, tr. D. A. Cress, 4th ed. Indianapolis: Hackett (original work published 1637). [Foundational works for modern philosophy, specifically the epistemological foundations of modern science and the (in)famous mind–body split. Undergraduates, graduates.]

Ess, C. 1996. "The political computer: democracy, CMC, and Habermas." In C. Ess, ed., *Philosophical Perspectives on Computer-mediated Communication*. Albany, NY: State University of New York Press, pp. 197–230. [An early effort to examine the democratization claims of CMC proponents in light of Habermas's theory of communicative reason. Undergraduates, graduates.]

——. 1999. "Critique in communication and philosophy: an emerging dialogue?" *Research in Philosophy and Technology* 18: 219–26. [A review of Chesebro & Bertelsen 1996 that examines their effort to conjoin philosophy and communication theory from a viewpoint primarily grounded in philosophy. Undergraduates, graduates.]

——. 2001. "What's culture got to do with it? Cultural collisions in the electronic global village, creative interferences, and the rise of culturally-mediated computing." Introduction, in C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 1–50. [Summarizes important cultural differences that become apparent in the implementation of Western CMC systems in diverse cultural settings, so as to argue that CMC technologies embed specific cultural values and communicative preferences, but that these can be resisted and overcome, especially through conscious attention to the social contexts of use, including education of CMC users. Undergraduates, graduates.]

Floridi, L. 1999. *Philosophy and Computing: An Introduction*. London and New York: Routledge. [Argues against the view that Cartesian philosophy – specifically dualism and the resulting mind–body split – plays the role often attributed to it in the development of and debates concerning Artificial Intelligence. Undergraduates, graduates.]

Friedman, T. 2001. "Cyber-serfdom." *New York Times*, Jan. 30, p. A27. [A prominent, technologically savvy, and attentive journalist's account of changing sentiments regarding information technology among some of the world's most influential business leaders. Undergraduates, graduates.]

Graham, G. 1999. *The Internet: A Philosophical Inquiry*. London and New York: Routledge. [One of the few book-length sustained examinations of the internet, notable for its somewhat more pessimistic conclusions, in contrast to the optimistic claims of enthusiasts, regarding the technology's promise as an agent of democratization and community. Undergraduates, graduates.]

Groleau, C. and Cooren, F. 1999. "A socio-semiotic approach to computerization: bridging the gap between ethnographers and systems analysts." *The Communication Review* 3(1,2): 125–64. [An example of using C. S. Peirce's theory of semiotics to increase the theoretical purchase of ethnography on HCI. Graduates.]

Hakken, D. 1999. *Cyborgs@cyberspace?: An Ethnographer Looks to the Future*. New York: Routledge. [An extended examination by a prominent ethnographer – one of the first to document the role of culture in CMC. Hakken establishes in multiple ways a more skeptical estimate of the future of CMC environments. Advanced undergraduates, graduates.]

Halavais, A. 2000. "National borders on the World Wide Web." *New Media and Society* 2(1): 7–28. [Quantitatively documents the correlation between culture and webpage production and consumption. Advanced undergraduates, graduates.]

Hamelink, C. 2000. *The Ethics of Cyberspace*. London: Sage. [A very readable summary of both theoretical discussion and pertinent research concerning multiple ethical issues – and at the same time a significant contribution towards more ethically-informed approaches to the design and implementation of CMC technologies, especially if they are to fulfill their promises of democratization (where democracy for Hamelink includes Habermasian dimensions). For both classroom and research use, undergraduate and graduate.]

Haraway, D. 1990. "A cyborg manifesto: science, technology, and socialist-feminism in the late twentieth century." In D. Haraway, ed., *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge, pp. 149–81. [A seminal essay in "cyber-feminism." While Haraway later modifies her views, this essay is still widely quoted and appealed to. Undergraduates, graduates.]

Harcourt. W., ed. 1999. *Women@internet: Creating New Cultures in Cyberspace*. London and New York: Zed Books. [Takes a global, interdisciplinary, and culturally oriented approach to women's engagement with the internet, in order to examine preexisting barriers to women's involvement and their multiple ways of reshaping the use of the technology to empower women both locally and globally. Undergraduates, graduates.]

Harris, R., Bala, P., Sonan, P., Guat Lien, E. K., and Trang, T. 2001. "Challenges and opportunities in introducing information and communication technologies to the Kelabit community of north central Borneo." *New Media and Society* 3(3) (Sept.): 271–96. [A research report on a project to implement internet connections within a remote village, significant for its efforts to do so by first establishing the values and priorities of the community and to involve the community in the implementation process so as to ensure that the technology and its uses are shaped by the community rather than vice versa. Undergraduates, graduates.]

Hayles, K. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press. [Centrally devoted to the questions of embodiment, with a solid grounding in the histories of science, technology, and culture; Hayles overcomes the usual Manichean dichotomies to establish a positive and promising middle ground between modernity and more extreme versions of postmodernism. Advanced undergraduates, graduates.]

Heaton, L. 2001. "Preserving communication context: virtual workspace and interpersonal space in Japanese CSCW." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 213–40. [Provides an overview of the (scant) CSCW literature that recognizes the role of culture, and documents important design projects in Japan that reflect distinctive communicative preferences and cultural values. Undergraduates, graduates.]

Herring, S. 2002. "Computer-mediated communication on the internet." In B. Cronin, ed., *Annual Review of Information Science and Technology*, vol. 36. Medford, NJ: Information Today Inc./ American Society for Information Science and Technology. [A masterful overview by one of the pioneers of CMC research, providing descriptions and definitions of diverse forms of CMC and a comprehensive discussion of research on communication in CMC environments as both similar to and distinctively different from previous modes and media of communication. Undergraduates, graduates.]

Hollan, J. D. 1999. "Human–computer interaction." In R. Wilson and F. Keil, eds., *The MIT Encyclopedia of the Cognitive Sciences*. Available online: <http://mitpress.mit.edu/MITECS/>. [Undergraduates, graduates.]

Hongladarom, S. 2000. "Negotiating the global and the local: how Thai culture co-opts the internet." *First Monday* 5(8) (July); <http://firstmonday.org/issues/issue5_8/hongladarom/index.html>. [A continuation of Hongladarom's analysis, focusing on the Thai online community of <www.pantip.com>. Undergraduates, graduates.]

——. 2001. "Global culture, local cultures and the internet: the Thai example." In C. Ess, ed.,

Culture, Technology, Communication: Towards an Intercultural Global Village. Albany, NY: State University of New York Press, pp. 307–24. [Documents how Thai users make use of CMC to preserve distinctive Thai cultural values, and provides a model (based on Michael Walzer) for both local and global uses of CMC that avoid cultural homogenization. Undergraduates, graduates.]

Hrachovec, H. 2001. "New kids on the net: Deutschsprachige Philosophie elektronisch." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 129–49. [Documents some of the earliest efforts to exploit CMC for the sake of philosophical discussion in German-speaking countries and specific cultural barriers encountered therein. Undergraduates, graduates.]

Human–Computer Interface Bibliography, <http://www.hcibib.org/>. [One of the most extensive online resources – noncommercial! – devoted to HCI. Undergraduates, graduates.]

Jones, S. 2001. "Understanding micropolis and compunity." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 51–66. [An excellent representation of postmodern approaches to CMC as well as an insightful and creative contribution to discussion of such basic issues as privacy, property, etc., and the prospects for online community. Online communities, Jones argues, are only partially successful, and they introduce in turn new difficulties distinctive to cyberspace. Undergraduates, graduates.]

Keniston, K. 2001. "Language, power, and software." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 281–306. [An extensive examination of both the need for and the multiple barriers to software localization in India – an important test case for the democratization claim, insofar as India is the world's largest democracy, and also the most diverse in terms of languages and cultures. Undergraduates, graduates.]

Kolb, D. 1996. "Discourse across links." In C. Ess, ed., *Philosophical Perspectives on Computer-mediated Communication*. Albany, NY: State University of New York Press, pp. 15–26. [One of the few philosophers who has published both in traditional print and hypertext, Kolb undertakes a nuanced philosophical inquiry into the matches and disparities between diverse argument styles and hypertext structures, one that balances enthusiasm for their possibilities with recognition of their limits, as well as the limits resulting from human beings as finite. Undergraduates, graduates.]

Landow, G. 1992. *Hypertext: The Convergence of Contemporary Critical Theory and Technology*. Baltimore: Johns Hopkins University Press. *Also*: Landow, G., ed. 1994. *Hyper/Text/Theory*. Baltimore: Johns Hopkins University Press. [Seminal texts by one of the most significant and influential theorists of hypertext. Undergraduates, graduates.]

Lyotard, J.-F. 1979 [1984]. *The Postmodern Condition: A Report on Knowledge*, tr. G. Bennington and B. Massumi. Minneapolis: University of Minnesota Press. [Perhaps the single most important foundation and springboard for postmodernism in the English-speaking world. Advanced undergraduates, graduates.]

Maitland, C. 2001. Personal communication.

Mayer, P. A. 1998. "Computer-mediated interactivity: a social semiotic perspective." *Convergence: The Journal of Research into New Media Technologies* 4(3): 40–58. [An example of using C. S. Peirce's semiotics as a way of analyzing CMC. Advanced undergraduates, graduates.]

McBeath, G. and Webb, S. A. 2000. "On the nature of future worlds?: considerations of virtuality and utopias." *Information, Communication & Society* 3(1): 1–16. [Connects contemporary discourse on CMC as creating utopian spaces with traditional utopian thinking and draws on Husserl (as opposed to what the authors see as the dominant Heideggerian approach) and Nozick to argue for designing software in specific ways in order to sustain defensible computer-mediated utopias. Advanced undergraduates, graduates.]

McChesney, R. 2000. "So much for the magic of technology and the free market: the world wide web and the corporate media system." In A. Herman and T. Swiss, eds., *The World Wide Web and Contemporary Cultural Theory*. New York and London: Routledge, pp. 5–35. [A highly critical evaluation of the role of commercial interests in shaping CMC technologies and their uses, by a prominent technology analyst. Undergraduates, graduates.]

Moravec, H. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press. [Moravec's optimistic

arguments and visions for the development of artificial intelligence and robots that will surpass, but not necessarily enslave, human beings. Undergraduates, graduates.]

Nietzsche, Friederich. [1954] 1988. "Twilight of the Idols" [*Götzen-Dämmerung*]. In *The Portable Nietzsche*, tr. W. Kaufmann. New York: Penguin, pp. 463–563. [One of Nietzsche's last works, this collection of aphorisms provides an unparalleled overview and self-evaluation of his work, its sources, and its hoped-for impacts – along with, for example, a number of historical and cultural observations from this self-proclaimed "untimely man." Undergraduates, graduates.]

O'Leary, S. and Brasher, B. 1996. "The unknown God of the internet: religious communication from the ancient agora to the virtual forum." In C. Ess, ed., *Philosophical Perspectives on Computer-Mediated Communication*. Albany, NY: State University of New York Press, pp. 223–69. [One of the first systematic overviews of "religion online," seen in part through the lens of rhetorical practice. While helpfully critical of "cyber-gnosticism" – a quest for "the information that saves" that fails to understand the difference between information and wisdom – the authors take an optimistic stance regarding the expansion of online spirituality and religious communities. Undergraduates, graduates.]

Ong, W. 1988. *Orality and Literacy: The Technologizing of the Word*. London: Routledge. [A key work in the development of the Eisenstein/Innis/McLuhan/Ong approach to communication, beginning with oral communication, as a form of technology – one that shapes fundamental assumptions regarding reason, the nature and role of argument *vis-à-vis* narrative, and the very structures (egalitarian vs. hierarchical) of societies. Undergraduates, graduates.]

Poster, M. 1997. "Cyberdemocracy: internet and the public sphere." In D. Porter, ed., *Internet Culture*. New York: Routledge, pp. 201–17. [One of the most significant analyses of the issues surrounding democracy and CMC from postmodern and feminist perspectives (including an important discussion and critique of Habermas's notion of the public sphere). Undergraduates, graduates.]

Plant, S. 2000. "On the matrix: cyberfeminist simulations." In G. Kirkup et al., eds., *The Gendered Cyborg*. London and New York: Routledge, pp. 265–75. [A complex article by one of the best-known writers on cyberfeminism, Plant takes up Irigarary's notion of the "specular economy" (that characterizes capitalism and patriarchy as a matter of trading women's bodies) and argues – somewhat parallel to early Haraway – that women can escape this enslavement in cyberspace. Plant is also of interest as she takes up Maturana's work in relation to cybernetics – thus continuing a discussion thread begun by Winograd and Flores. Advanced undergraduates, graduates.]

Richards, C. 1998. "Computer mediated communication and the connection between virtual utopias and actual realities." In C. Ess and F. Sudweeks, eds., *Proceedings of the First International Conference on Cultural Attitudes Towards Technology and Communication*. Sydney: Key Centre Design Computing, pp. 129–40. [One of the (relatively) early analyses of the utopian/dystopian dichotomy characteristic of much, especially postmodern, discourse concerning CMC that takes up a phenomenological framework (Ricoeur) in order to find a more balanced view of the utopian possibilities of CMC. Undergraduates, graduates.]

Sandbothe, M. 1999. "Media temporalities of the internet: philosophies of time and media in Derrida and Rorty." *AI and Society* 13(4): 421–34. [Sandbothe summarizes diverse philosophies of time in the twentieth century, and argues that the experience of time in CMC environments favors constructivist views – views that are further consonant with Derrida and Rorty in arguing for a decentered, pluralist world of mutual understanding rather than dominance by a single culture and its values. Advanced undergraduates, graduates.]

Sardar, Z. 2000. "ALT.CIVILIZATIONS.FAQ: cyberspace as the darker side of the west." In D. Bell and B. Kennedy, eds., *The Cybercultures Reader*. London and New York: Routledge, pp. 732–52. [An exceptionally powerful critique of the overt and covert forms of colonialism the author sees in the technologies and discourses surrounding cyberspace. Undergraduates, graduates.]

Scaltsas, T. 1998. "Representation of philosophical argumentation." In T. Bynum and J. Moor, eds., *The Digital Phoenix: How Computers are Changing Philosophy*. Oxford: Blackwell, pp. 79–92. [An account of the Archelogos Project, which uses computers to help uncover and visually represent argument structures in ancient Greek texts. Undergraduates, graduates.]

Shank, G. and Cunningham, D. 1996. "Mediated phosphor dots: toward a post-Cartesian model of CMC via the semiotic superhighway." In C. Ess, ed., *Philosophical Perspectives on Computer-mediated Communication*. Albany, NY: State University of New York Press, pp. 27–41. [Using C. S. Peirce's notion of semiotics, the authors critique prevailing distinctions between orality and textuality and argue that a semiotic conception of self and communication may lead to an "age of meaning" rather than a putative Information Age. Undergraduates, graduates.]

Stevenson, N. 2000. "The future of public media cultures: cosmopolitan democracy and ambivalence." *Information, Communication & Society* 3(2): 192–214. [A theoretically and practically informed analysis of the conditions required for a cosmopolitan democracy in a global media culture, concluding with policy recommendations for fostering the author's "cautious cosmopolitanism." Undergraduates, graduates.]

Stratton, J. 1997. "Cyberspace and the globalization of culture." In D. Porter, ed., *Internet Culture*. New York and London: Routledge, pp. 253–75. [A helpful critique of the assumptions underlying American optimism regarding the internet – including a misleading nostalgia for lost community, and the presumption that American values (human rights, individualism, and democracy – coupled with capitalism) are rightly "the homogenizing basis of the internet community" (271). Stratton argues that such globalizing/homogenizing bases, however, threaten to exclude a plurality of cultures, peoples, and non-economic interests. Undergraduates, graduates.]

Suchman, L. 1987. *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge and New York: Cambridge University Press. [Argues against a European view of planning and action as exclusively rational and abstract – a view further at work in the design of intelligent machines at the time – in favor of an account of purposeful actions as situated, i.e., actions "taken in the context of particular, concrete circumstances" (p. viii), as a more fruitful basis for design of computers and HCI. Advanced undergraduates, graduates.]

Sy, P. 2001. "Barangays of IT: Filipinizing mediated communication and digital power." *New Media and Society* 3(3) (Sept.): 297–313. [Draws on Habermas and Borgmann to argue for a "cyber-barangay" – in part, as a synthesis of the oral and the textual – as a possible alternative to the spread of CMC as otherwise a colonization of the lifeworld. Undergraduates, graduates.]

Wertheim, M. 1999. *The Pearly Gates of Cyberspace: A History of Space from Dante to the Internet*. New York: Norton. [By developing an account of space from the Middle Ages through modern science and cyberspace, Wertheim argues strongly against the claims that cyberspace may hold transcendental or redemptive potentials. Undergraduates, graduates.]

Wheeler, D. 2001. "New technologies, old culture: a look at women, gender, and the internet in Kuwait." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 187–212. [One of the very few analyses of the impacts of CMC among women in the Muslim world. Wheeler finds that CMC partially fulfills the democratization promise of its Western proponents, thereby changing the social patterns of interaction between men and women in Kuwait. Undergraduates, graduates.]

Willis, A. 2000. "Nerdy no more: a case study of early *Wired* (1993–96)." In F. Sudweeks and C. Ess, eds., *Second International Conference on Cultural Attitudes Towards Technology and Communication 2000*. Murdoch, Australia: School of Information Technology, Murdoch University, pp. 361–72. Available online: <http://www.it.murdoch.edu.au/~sudweeks/catac00/> [A detailed analysis of *Wired* magazine, arguing that, contrary to the magazine's ostensibly democratic/egalitarian ideology, the prevailing content and images reinforce the power and viewpoints of upper-class white males. Undergraduates, graduates.]

Wilson, R. A. and Keil, F., eds. 1999. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press. Available online (subscription fee for full-text access): <http://mitpress.mit.edu/MITECS/>. [An invaluable resource, made up of both general summary articles and highly detailed discussions and analyses of specific issues, topics, and figures in cognitive science. Undergraduates, graduates.]

Winograd, T. and Flores, F. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Reading, MA: Addison-Wesley. [A seminal effort to conjoin biologically based learning theory with philosophical epistemology in order to articulate the philosophical foundations for

Human–Computer Interface design. Advanced undergraduates, graduates.]

Yoon, S. 1996. "Power online: a poststructuralist perspective on CMC." In C. Ess, ed., *Philosophical Perspectives on Computer-mediated Communication*. Albany, NY: State University of New York Press, pp. 171–96. [Takes up Foucault's notion of positive power to analyze ways in which CMC in Korea work in antidemocratic directions. Undergraduates, graduates.]

——. 2001. "Internet discourse and the habitus of Korea's new generation." In C. Ess, ed., *Culture, Technology, Communication: Towards an Intercultural Global Village*. Albany, NY: State University of New York Press, pp. 241–60. [Using both theoretical approaches (Bourdieu, Foucault) and ethnographic interviews to examine the role of Korean journalism and the function of such social dynamics as "power distance" (Hofstede) to document the antidemocratic impacts of CMC as it extends into Korean youth culture, especially as these are fostered by commercialization. Undergraduates, graduates.]

Chapter 7

Internet Culture

Wesley Cooper

Introduction

The internet is a magnet for many metaphors. It is cyberspace or the matrix, the "information superhighway" or infobahn or information hairball, a looking-glass its users step through to meet others, a cosmopolitan city with tony and shady neighborhoods, a web that can withstand nuclear attack, electric Gaia or God, The World Wide Wait, connective tissue knitting us into a group mind, an organism or "vivisystem," a petri dish for viruses, high seas for information pirates, a battleground for a war between encrypters and decrypters, eye candy for discreet consumers of a tsunami of pornography, a haven for vilified minorities and those who seek escape from stultifying real-world locales, a world encyclopedia or messy library or textbook or post office, chat "rooms" and schoolrooms and academic conferences, a vast playground or an office complex, a cash cow for the dot.coms, The Widow Maker, training wheels for new forms of delinquency practiced by script kiddies and warez d00des, a wild frontier with very little law and order, the glimmer in the eyes of virtual-reality creators, a workshop for Open Source programmers, a polling booth for the twenty-first century, a marketplace for mass speech, a jungle where children are prey, a public square or global village, a mall or concert hall, a stake for homesteaders, a safari for surfers, a commercial space much in need of zoning, the mother of all Swiss Army knives, a tool palette for artists, a lucid dream or magic, a telephone or newspaper or holodeck, a monster that has escaped DARPA's control, the Linux penguin, sliced bread, an addiction, the Grand Canyon, and on and on.

Before attempting to think through these metaphors, it is worthwhile to note at the outset that we regular users of the internet are only a minority, even in societies that have passed through industrialization and are now exploring economies in which information technology has become central. There are important technical, moral, and political issues about conversion of this minority into a majority, including whether that would be a good thing, whether it is required by fairness, how much priority should be given to information technology in developing countries especially relative to processes of industrialization, and so forth. It is clear, however, that desire for connection to the Net is not a minority taste, something only for a military or academic elite, but rather it corresponds closely to the enormous demand for the ubiquitous computer itself at every social level. So the prospect of a global electronic metropolis, in which citizens can reliably be expected to be netizens, is not an idle dream, or nightmare. The internet is so new that we don't know yet whether it has an Aristotelian *telos* of some benign or malign nature, or

whether instead it will always be a loose and disjointed Humean thing, evading every attempt to discern an underlying unity.

Although the internet is bringing us together, it also keeps us apart in two general ways. First, time spent online is inevitably time spent in a greater or lesser degree of detachment from one's physical surroundings, including local others. Second, the connection to distant others is itself a form of detachment, as coolly a matter of business as online banking or as etiolated a form of sociability as a chat room. The major issue about the former is simply time management. Almost everyone has decided that local detachment is all right, because we do it when reading, watching television, listening to music, and so forth. But there are still questions about how internet use will impact on these other forms of detachment, for instance in reading less or, worse, less well. The latter issue is more complex. Detachment from distant others can be valued for purposes of efficiency, as with banking, or because it affords anonymity to members of unpopular subcultures, as with some chat rooms, or because one happens to find that level of sociability to one's liking. There is not anything evidently wrong with any of this, putting criminal or pathological cases aside – hacking into banks, planning terrorist attacks, escaping from life, and so on. Perhaps the major issue about online detachment will have to do with its transformation as more "bandwidth" gets piped into our ever more versatile computers, giving them audiovisual and even tactile powers to create experiences that are very different from invoking File Transfer Protocol from a command line to send scientific data from node A to node B.

The internet is also changing us. Users of the internet are not the people they would have been in the absence of the computer revolution. At one level this is a truism: experiences change us. But many interpreters of postmodern culture, the culture of postindustrial societies particularly as influenced by information technology such as the internet (and computers, CDs, etc.), detect a change in us that is understated even by emphasizing that our personalities have become different. Some of these interpretations are pretentious babble, including much theorizing that passes as postmodernist philosophy or psychology when it opines that there is nothing outside the text, that the self is an outmoded social construct, and so forth. Postmodernist theory should be sharply distinguished from postmodern culture. The latter, however it is to be characterized in detail, is a large social fact; the former, whether it is true or false or meaningful or nonsensical, is precisely a theory; one can be a participant in postmodern culture without espousing postmodernist doctrine. Interpretations of postmodern culture, including many insightful ones, point to the need for a theory of personhood and personal identity that does full justice to the changes in us, and gives us a way of thinking constructively about them. Many different disciplines, from philosophy to psychology, from linguistics to sociology, from anthropology to literary studies, should converge so as to develop such a theory.

The internet is changing our relationship to nature, not only in the way that postmodernist theorists emphasize, by "thickening" the layers of images that mediate our perception of the external world and our interactions with it, but also by starting to lessen the stress on nature caused by the technologies of the industrial revolution. The two are related. The thickened layers can include the images that constitute the emerging technology of teleconferencing, and the lessened stress, we have reason to hope, will take the form of reduced environmental damage caused by planes, trains, and automobiles; alternatives to fossil fuel will depend, either at the research stage or in implementation, on digital technology to harness the energy of the sun, the wind, hydrogen, and so forth. The layers can include the electronic paper that is clearly visible now on the technological horizon, and the relief for nature will be felt by our forests. The power of computer modeling should also be mentioned, a new way of representing the world that is proving its value for understanding, monitoring, and controlling natural processes, from the human genome to the weather; it is changing the way traditional sciences are undertaken as well as birthing relatively new sciences such as cognitive psychology, artificial intelligence, and nanotechnology. These changes in the images or representations that we rely upon are introducing social changes as well, ranging from less reliance on the amenities of cities for educational and entertainment purposes, to new forms of populism as

groups organize on the internet despite lack of access to high-cost tools. The great engine of acculturation, schooling, is now producing generations for whom computer use is second nature, a presence in the classroom since the first year. This large fact presents a challenge to the existence of a "mainstream culture," since this generation will be influenced by such various cultural forces that even the cultural fragmentation occasioned by the 500-channel television will only hint at the upshot. Let this thought be the background to the question whether the internet and information technology not only are having impact on the larger culture, but also whether they have a culture of their own.

Internet Culture?

Is there internet culture, something more substantial than shared mastery of the email or chatroom "smiley," or is that an oxymoron? Is the internet a tool, or something more? Is the internet improving education or corrupting it? Is the space of cyberspace a place to explore utopian possibilities, or a wrecking yard for traditional culture, or something as neutral with respect to questions of value as a screwdriver? These are some of the questions that a philosophy of internet culture should address. The answers to be found in a large and diverse literature on the subject are classifiable as utopian, dystopian, or instrumental. A utopian view sees the internet as good, perhaps profoundly so, or at least good-on-balance. As dystopian, it is profoundly bad or at least bad-on-balance. And as instrumental, the Net is a tool, perhaps merely a tool or at least a tool that does not harbor profoundly good or evil values.

The notion of profundity in this trichotomy acknowledges the influence of Martin Heidegger on the philosophy of technology, especially his *The Question Concerning Technology*. Many interpreters of the internet have borrowed from him the idea that a technology can be inseparable from a value commitment. Heidegger would not have liked the term "value." In "Letter on Humanism" he writes, "Every valuing, even where it values positively, is subjectivising. It does not let beings: be . . . the thinking that inquires into the truth of Being and so defines man's essential abode from Being and toward Being is neither ethics nor ontology" (1977: 87). This chapter returns to Heidegger under the heading of dystopian inherence, making the case that Heidegger's philosophy of technology does indeed betray a significant value commitment, contrary to its aim at something more profound, a commitment that undermines its authority as a model for understanding the internet.

The general *Heideggerian* idea of a value inherent in technology is instanced in the statement that the high technology of factory farming, or "agribusiness," is inseparable from a bad way of relating to nature, understanding it, and treating it simply as something to be processed in wholesale fashion for satisfaction of human appetites. Heidegger's idea has been adopted mainly by dystopian theorists like his translator Michael Heim, who argues in *The Metaphysics of Virtual Reality* that the "Boolean logic" of the computer marks a "new psychic framework" that "cuts off the peripheral vision of the mind's eye" and generates infomania (1993: 22, 25), as he indicates in the following passage.

> Note already one telltale sign of infomania: the priority of system. When system precedes relevance, the way becomes clear for the primacy of information. For it to become manipulable and transmissible as information, knowledge must first be reduced to homogenized units. With the influx of homogenized bits of information, the sense of overall significance dwindles. This subtle emptying of meaning appears in the Venn diagrams that graphically display Boolean logic. (1993: 17)

Heim's profound or *inherence* dystopianism may be contrasted with on-balance or simply *balance* dystopianism, exemplified by Sven Birkerts' *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*, and particularly by the cost–benefit analysis of the computer revolution that he provides in the following passage.

> We can think of the matter in terms of gains and losses. The gains of electronic postmodernity could be said to include, for individuals,

> (a) an increased awareness of the "big picture," a global perspective that admits the extraordinary complexity of interrelations; (b) an expanded neural capacity, an ability to accommodate a broad range of stimuli simultaneously; (c) a relativistic comprehension of situations that promotes the erosion of old biases and often expresses itself as tolerance; and (d) a matter-of-fact and unencumbered sort of readiness, a willingness to try new situations and arrangements.
>
> In the loss column, meanwhile, are (a) a fragmented sense of time and a loss of the so-called duration of experience, that depth phenomenon we associate with reverie; (b) a reduced attention span and a general impatience with sustained inquiry; (c) a shattered faith in institutions and in the explanatory narratives that formerly gave shape to subjective experience; (d) a divorce from the past, from a vital sense of history as a cumulative or organic process; (e) an estrangement from geographic place and community; and (f) an absence of any strong vision of a personal or collective future. (Birkerts 1994: 27)

Note that the distinction between inherence and balance dystopians concerns the form of argumentation rather than conclusions about the technology, which may be similar. Heim would agree with Birkerts that, as the latter writes, "We are at a watershed point. One way of processing information is yielding to another. Bound up with each is a huge array of aptitudes, assumptions, and understandings about the world" (1994: 27). But Heim has an extra reason for that conclusion, the profound one about the "infomania" value inherent in the new technology.

Heidegger's idea, this extra reason, can be extended to utopianism. An inherence utopian about the internet, on this extension, is one who believes that there is something good about it beyond a simple toting up of gains and losses. For instance, *Wired* magazine editor Kevin Kelly's *Out Of Control: The New Biology of Machines, Social Systems, and the Economic World* theorizes the internet as a *vivisystem*, and as such an instance of, in his words,

> [t]he overlap of the mechanical and the lifelike [that] increases year by year. Part of this bionic convergence is a matter of words. The meanings of "mechanical" and "life" are both stretching until all complicated things can be perceived as machines, and all self-sustaining machines can be perceived as alive. Yet beyond semantics, two concrete trends are happening: (1) Human-made things are behaving more lifelike, and (2) Life is becoming more engineered. The apparent veil between the organic and the manufactured has crumpled to reveal that the two really are, and have always been, of one being. What should we call that common soul between the organic communities we know of as organisms and ecologies, and their manufactured counterparts of robots, corporations, economies, and computer circuits? I call those examples, both made and born, "vivisystems" for the lifelikeness each kind of system holds. (1994: 3)

The inherent value for Kelly is the value of a vivisystem, as revelatory of a hidden connection between the natural and the mechanical. Kelly's focus on vivisystems is comparable to historian Bruce Mazlish's reconstruction of how we have overcome the fourth discontinuity, between ourselves and machines, the earlier discontinuities having been overcome when Copernicus showed that our earth was not the center of the universe, when Darwin showed that man did not have a privileged place in creation, and when Freud showed that our rationality is not so perfect as to set us apart from the other animals. Kelly's vivisystems allow Mazlish's point to be put positively, in terms of continuity rather than discontinuity: the range of manmade and natural vivisystems reveals the continuity between ourselves and machines.

Vivisystems figure in the version of James Lovelock's *Gaia* Hypothesis that Kelly endorses. This is the hypothesis, that, in Lovelock's words, "The entire range of living matter on Earth, from whales to viruses, from oaks to algae, could be regarded as constituting a single living entity, capable of manipulating the Earth's atmosphere to suit its overall needs and endowed with faculties and powers far beyond those of its constituent parts" (Kelly 1994: 83). (Kelly is quoting from Lovelock's *The Ages of Gaia*.) Although there may be controversy about whether Gaia is an organism, Kelly thinks there should be no

doubt that, as he writes, "it *really is* a system that has living characteristics. It is a vivisystem. It is a system that is alive, whether or not it possesses all the attributes needed for an organism" (1994: 84). Gaia is not only alive but it is coming to have a mind, thanks to the internet and other networking technologies. Kelly makes the point in dramatic language.

> There is a sense in which a global mind also emerges in a network culture. The global mind is the union of computer and nature – of telephones and human brains and more. It is a very large complexity of indeterminate shape governed by an invisible hand of its own. We humans will be unconscious of what the global mind ponders. This is not because we are not smart enough, but because the design of a mind does not allow the parts to understand the whole. The particular thoughts of the global mind – and its subsequent actions – will be out of our control and beyond our understanding. Thus network economics will breed a new spiritualism.
>
> Our primary difficulty in comprehending the global mind of a network culture will be that it does not have a central "I" to appeal to. No headquarters, no head. That will be most exasperating and discouraging. In the past, adventurous men have sought the holy grail, or the source of the Nile, or Prester John, or the secrets of the pyramids. In the future the quest will be to find the "I am" of the global mind, the source of its coherence. Many souls will lose all they have searching for it – and many will be the theories of where the global mind's "I am" hides. But it will be a never-ending quest like the others before it. (1994: 202)

Another inherence-utopian vision incorporates the internet's group mind as only a minor foreshadowing of an end-of-time God, intelligent life connected throughout the universe, as a result of colonization of space (and so forth). It will tap into the energy created by gravity's "divergence towards infinity" in the Big Crunch so as to reproduce all past experience in massive computations that generate the requisite virtual realities. Construing our brains as virtual reality generators themselves, these theorists prophesy that brains can be replaced by their Turing-machine essence: we will be brought back to life as programs suitable for generating the virtual-reality renderings that capture our lived experience, with the unpleasant bits trimmed away and desirable additions inserted, perhaps additions from program-based future societies, if we can tolerate the culture shock. The details can be found in Frank J. Tipler's *The Physics of Immortality* and David Deutsch's *The Fabric of Reality*.

This much will serve to introduce a framework for understanding internet culture and the theorizing that surrounds it: the utopian/dystopian/instrumental trichotomy and the balance/inherence dichotomy. The stage is set for a critical illustration of balance utopianism, in the next section; then inherence dystopianism; and then inherence instrumentalism; and finally some concluding remarks, including some caveats and qualifications about the framework just bruited.

Balance Utopianism

The advent of the internet took Sherry Turkle by surprise. She had published *The Second Self* in 1984, describing the identity-transforming power of the computer at that stage of the computer revolution. Reflecting on her experience and the experience of others with the new Apple and IBM PC computers, she conceived of the relationship of a person to her computer as one-on-one, a person alone with a machine. By 1995, when *Life on the Screen* appeared, she was writing about something quite different, "a rapidly expanding system of networks, collectively known as the internet, [which] links millions of people in new spaces that are changing the way we think, the nature of our sexuality, the form of our communities, our very identities" (1995: 9).

Though Turkle speaks neutrally here of "change" in these matters, she fits into the "utopian" category of her trichotomy between utopian, apocalyptic, and utilitarian evaluations of the internet. The computer is a new and important tool, most assuredly, but the internet makes it "even more than a tool and mirror: We are able to step through the looking glass. We

are learning to live in virtual worlds. We may find ourselves alone as we navigate virtual oceans, unravel virtual mysteries, and engineer virtual skyscrapers. But increasingly, when we step through the looking glass, other people are there as well" (1995: 9). Whereas apocalyptic theorists diagnose this as stepping through the looking-glass to cultural impoverishment or a new form of mental illness, Turkle theorizes the new experiences by reference to colonization of a new land.

This metaphor of colonization should be understood carefully, however, as she is not suggesting that Sherry Turkle, sociologist and MIT professor, should be left behind in favor of a new life as the cybernaut ST on LambdaMOO. That suggestion comes from an extreme form of inherence utopianism about the internet, or it is the equally extreme suggestion of inherence dystopian theorists, like Mark Slouka in *War of the Worlds*, who diagnose the internet experience as equivalent to wholescale departure from everyday reality. More in Turkle's spirit is the thought that a new dimension of human life is being colonized, and although that raises a host of new issues about budgeting time and effort, and even about physical and mental health, Turkle is not proposing that it be undertaken in the spirit of these extreme forms of utopianism.

She does indeed characterize her colonists as "constructing identity in the culture of simulation," in a cultural context of "eroding boundaries between the real and the virtual, the animate and the inanimate, the unitary and the multiple self" (1995: 10), a context in which experiences on the internet figure prominently but share a cultural drift with changes in art, such as the postmodern architecture that the cultural critic Fredric Jameson studies; science, such as research in psychoanalysis and elsewhere inspired by connectionist models of the mind/brain; and entertainment, such as films and music videos in which traditional narrative structure is hard to discern. Constructing identity in the culture of simulation – our postmodern culture, as Turkle interprets it – involves two closely related ideas. First, there is the idea that we are newly aware of a rich continuum of states between the real and the virtual, the animate and the inanimate, the unitary and the multiple self. A boundary that may have been a sharp line is now a complex zone. For instance, a player who manipulates a character or avatar in an online virtual reality such as a Multi-User Dungeon (MUD) is distinctly located in that zone. By contrast, traveling to Rome or viewing someone's movie about Rome, even when doing so is "virtually like being there," is safely on one side or the other of the real/virtual line, awakening no awareness of the zone being constructed and explored by Turkle's colonists.

Second, constructing identity involves something like the notion of a dimension as it was just introduced: although Turkle is distinctly on the "real" side of the real/virtual continuum, she now builds her identity partially by reference to dimensions of herself that owe their existence to activity in the border zone. To the degree that MUDing is important to her, for instance, to that degree it is constitutive of who she is. This is a high-technology application of the general principle that we are self-defining creatures. It is *not* the idea that crossing the postmodern divide has somehow destroyed personal identity. Although some psychologists and sociologists adopt the conceit of speaking this way, it is no more than acknowledging the complexity of self-definition in modern society; or else this way of speaking falsely equates personal identity with a soul-pellet or Cartesian Thinking Substance, in which case it is broadcasting the stale news that such conceptions of the self are largely discredited. Turkle discusses the phenomenon of Multiple Personality Disorder, and it may be that MPD is more common because of the stresses of modern life, and not because, say, the medicalization of human experience leads us to find mental illnesses today that weren't there yesterday. But constructing identity is, and always has been, distinct from going crazy, even when the building material is a new high-tech dimension.

This is not to say that Turkle always gets this exactly right. Setting out some of her interviews with students who play MUDs, she writes that "as players participate, they become authors not only of text but of themselves, constructing new selves through social interaction. One player says, 'You are the character and you are not the character, both at the same time.' Another says, 'You are who you pretend to be.'" Analyzing these interviews, she continues, "MUDs make possible

the creation of an identity so fluid and multiple that it strains the limits of the notion. Identity, after all, refers to the sameness between two qualities, in this case between a person and his or her persona. But in MUDs one can be many" (1995: 12). The short path out of these woods is to deny that a person and his or her persona are identical: you are not who you pretend to be, but rather you are pretending to be someone in such a way as to call upon your verbal, emotional, and imaginative resources to accomplish the pretense.

One of Turkle's major themes is the transition from modern to postmodern culture, which she glosses as follows, beginning with a set of ideas that have come to be known as "postmodernism."

> These ideas are difficult to define simply, but they are characterized by such terms as "decentered," "fluid," "nonlinear," and "opaque." They contrast with modernism, the classical world-view that has dominated Western thinking since the Enlightenment. The modernist view of reality is characterized by such terms as "linear," "logical," "hierarchical," and by having "depths" that can be plumbed and understood. MUDs offer an experience of the abstract postmodern ideas that had intrigued yet confused me during my intellectual coming of age. In this, MUDs exemplify a phenomenon we shall meet often in these pages, that of computer-mediated experiences bringing philosophy down to earth. (1995: 17)

It does so, Turkle suggests, because the transition from modernism to post-modernism, from the early post-Second World War years onward, is paralleled in the world of computers by a transition from a culture of calculation to a culture of simulation. For those caught up in the war effort, like John von Neumann, the new computers were objects to calculate with, specifically to make the staggeringly complex calculations that would tell whether an implosion device would detonate an atomic bomb. Even the relatively carefree hackers at the MIT AI Lab in the fifties and sixties were privy to this culture, prizing what Turkle calls "vertical" understanding of the computer: understanding it all the way down from high-level programming languages to assembler to machine language, and wanting to know as well the engineering architecture of the hardware. (Hackers who loved to code but knew little about hardware were called "softies.") By contrast the consumer computers that were brought to the market in the mid-seventies to early eighties, first by Apple and then by IBM and many others, made computers accessible far beyond the military, industry, and academe. For Turkle the Apple Macintosh's graphical user interface, as well as its presenting itself as "opposed and even hostile to the traditional modernist expectation that one could take a technology, open the hood, and see inside" (1995: 35), are crucial developments, giving the computer massive popular appeal to many who preferred "horizontal understanding," of an operating system's or an application's interface, surface over depth.

> The power of the Macintosh was how its attractive simulations and screen icons helped organize an unambiguous access to programs and data. The user was presented with a scintillating surface on which to float, skim, and play. There was nowhere visible to dive. (1995: 34)

The massive growth of internet culture, from its roots in the MIT/ARPANET connection and the UNIX/USENET connection, into the behemoth we see now, turned on the fact that a lot of people want to be pilots, not mechanics.

Turkle acknowledges that even her beloved Macintosh ultimately requires the skills and tools of modernist culture, but it strove to make these "irrelevant" to the user, and in this way "the tools of the modernist culture of calculation became layered underneath the experience of the culture of simulation" (1995: 34). This is an important point, and one that she may not have developed sufficiently. *The culture of simulation requires a modernist spine*. It requires technicians to keep its computer network running, for one thing, but it also needs inventors and theoreticians to explore its possibilities. More generally, it needs a background of a world that is external to its rapidly thickening layers of images and other representations, a world that is best disclosed by

the sciences, in contradistinction to the postmodern conceit that there is nothing outside the text, that science is just one among many narratives in an anarchic cacophony, etc. Often enough to counsel attention, modernist values consort with plain truths. (This of course rejects the postmodernist theoretician's notion that truth reduces to what passes for true, which is a function of which community's values you subscribe to.) The plain truth of science's superior track record consorts with the modernist value that discerns a hierarchy in which science ranks higher than, say, wishful thinking in its power to reveal the nature of things. The plain truth that there is an external world consorts with the modernist value of depth, in this case a depth beyond our images, symbols, and other representations. The modernist value of prudence, of rational self-interest which gives equal weight to each moment of one's life, consorts with the plain truth about personal identity that I canvassed earlier. The value and the fact are not the same: one can grant that there is personal identity through time and rational concern about it, without embracing the modernist conception of prudence that requires one to be a shepherd, so to speak, for a whole human life. For instance, it is not irrational, on certain conceptions of rationality, to severely discount one's distant future. But such conceptions aren't those that have had influence in building our senses of ourselves and our social institutions, like social and medical insurance. Those reflect modernist values.

Inherence Dystopianism

A leitmotiv of some dystopian critique is a fallacy: an inference from features of computation to features of the media that the computation enables. Call this the Frame Fallacy, after the mistake of inferring from the fact that a movie is made up of discrete frames, the conclusion that the experience of watching a movie is the experience of a series of discrete frames.

For instance, Fred Evans makes observations about the algorithmic character of computation and infers from this that computer scientists and cognitive psychologists are in league with technocratic bureaucrats who are concerned only with efficient administration. There are in fact two fallacies here. First, efficient administration with respect to programming might be put to the service of organizations that are devoted to human rights and opposed to technocratic manipulation of citizens. To suppose the contrary to is to commit the simple Frame Fallacy. Additionally, Evans makes an unwitting philosophical pun – a fallacy of equivocation – on the term *efficiency*. The two fallacies blend in a spectacular howler.

Evans's *Psychology and Nihilism: A Genealogical Critique of the Computational Model of Mind* argues that "technocratic rationality" is a secret value presupposed by the computer model of mind, which he takes to be the model that defines cognitive science and cognitive psychology. His fear ("the crisis of modernity") is that consciousness itself "might be reduced to just those parameters necessary for the continued reproduction of restrictive and univocal social, cultural, and economic systems" (1993: 2). In this way the computer model of cognitive psychology "serves the interest of the new technocratic elite by emulating their style of thinking" (1993: 7). Assimilating us to machines, cognitive psychology implicitly denies those cultural values that affirm and celebrate life, and consequently it is "nihilist."

Evans's main argument is as follows.

> Because we can precisely state its properties, we shall use the Turing machine as our formalization and the idealization of "analytic discourse." Like analytic discourse, the Turing machine divides its subject matter into a set of discrete entities, maintains a strict separation between its program (language) and the domain over which it operates (the same program can imitate many different machines), adheres to an ideal of transparency in its code and in what it codifies, and subordinates its subject matter to the achievement of a preestablished goal that requires no change in the basic rules and symbols of the Turing machine's own program (the ideal of "domination" or "administration"). *For both analytic discourse and the Turing machine, the ideal is to transform everything into an "effective procedure," and this is exactly the task of technocratic rationality.* In more historical terms, the Turing

> machine transforms the "clearness and distinctness" dictum of Descartes into "imitatable by the Turing machine." (1993: 64)

At bottom, this argument is a bad pun. Evans is equivocating on "effective procedure," between *cost-effective administration* on one hand, and *algorithm* on the other. It is the same sort of mistake as supposing that, since the Bank of Montreal and the bank of the Saskatchewan River are both *banks*, it must follow that they both make financial transactions. Effective procedures in the sense that interested Alan Turing are features of mathematical reasoning, not features of administration of people. Evans's mistaken inference from features of computation to features of the research communities that make use of them is egregiously abetted by his equivocation on "effective procedure."

One reason to be wary of utopian or dystopian inherence theories is that they encourage a tendency toward blanket denunciation and renunciation of the internet, or the blanket opposite, when what is needed is a piecemeal evaluation of this or that use of it, this or that tool that is enabled by the internet metatool. A striking contemporary instance of the blanket approach is the Montana philosopher Albert Borgmann's position, in *Holding on to Reality*, that digitally generated information is incapable of making a positive contribution to culture, but on the contrary threatens to dissolve it, by introducing *information as reality* to compete with the picture of the world that is drawn from natural information (information *about* reality, as in weather reports) and cultural information (information *for* reality, as in recipes for baking things).

> The technological information on a compact disc is so detailed and controlled that it addresses us virtually *as* reality. What comes from a recording of a Bach cantata on a CD is not a report about the cantata nor a recipe – the score – for performing the cantata, it is in the common understanding music itself. Information through the power of technology steps forward as a rival of reality.
>
> Today the three kinds of information are layered over one another in one place, grind against each other in a second place, and are heaved and folded up in a third. But clearly technological information is the most prominent layer of the contemporary cultural landscape, and increasingly it is more of a flood than a layer, a deluge that threatens to erode, suspend, and dissolve its predecessors. (1999: 2)

This has led some disciples of Borgmann to eschew all digitally recorded music, insisting on listening only to live performances. Another example of inherence dystopianism leading to blanket evaluations is Neil Postman's *Technopoly: The Surrender of Culture to Technology*, which indicts the United States as a "technopoly," along with "Japan and several European nations that are striving to become Technopolies as well" (1993: 48–9). A Technopoly does no less, according to Postman, than eliminate "alternatives to itself in precisely the way Aldous Huxley outlined in *Brave New World*" (1993: 48).

An object lesson about the wholesale approach can be drawn from Richard Bernstein's analysis of the father of dystopian theories of high technology, Heidegger. In "Heidegger's Silence?: *Ethos* and Technology" Bernstein makes the case that the great German philosopher's brief but active support of Hitler and the Nazis, during the 10-month period when he served as Rector of the University of Freiburg between April 1933 and February 1934, is symptomatic of a *philosophical* failing that expresses itself in what he said and did before and after those 10 months, notably in his silence about the Holocaust after the war, when there were no longer any serious doubts about the full horror of the Nazi regime. "But we are delivered over to [technology] in the worst possible way," Heidegger writes in *The Question Concerning Technology*, "when we regard it as something neutral; for this conception of it, to which today we particularly like to do homage, makes us blind to the essence of technology" (1977: 91).

According to Bernstein's account of the link between his biography and his philosophy, Heidegger conceals and passes over in silence the importance for the Greeks, specifically Aristotle, of *phronesis*, the state of the soul that pertains to *praxis*. He refers to a discussion by Aristotle in *Nicomachean Ethics* that has "special

importance," but his reference is partial and one-sided, bringing out the role of *techne* in relation to *poiesis*, as sketched above, but not tracking the full discussion, which Aristotle summarizes in the following passage.

> Then let us begin over again, and discuss these states of the soul. Let us say, then, that there are five states in which the soul grasps the truth [*aletheia*] in its affirmations or denials. These are craft [*techne*], scientific knowledge [*episteme*], [practical] intelligence [*phronesis*], wisdom [*sophia*], and understanding [*nous*] . . . (cited in Bernstein 1992: 121)

Bernstein asks, "Why should we think that the response that modern technology calls forth is to be found by "re-turning" to *techne* and *poiesis*, rather than *phronesis* and *praxis*?" He objects that Heidegger does not even consider this possibility, writing that "[t]he entire rhetorical construction of *The Question Concerning Technology* seduces us into thinking that the only alternative to the threatening danger of *Gestell* is *poiesis*. It excludes and conceals the possibility of *phronesis* and *praxis*" (1992: 122). Bernstein urges that our destiny rests not solely with the thinkers and the poets who are guardians of the abode in which man dwells, but with the *phronesis* of ordinary citizens' contribution to public life. The possible upsurgence of the saving power may be revealed in action (*praxis*) and not only in "poetic dwelling."

Bernstein asks again, "Why is Heidegger blind to those aspects of *praxis* and *phronesis* highlighted by Taminiaux, Gadamer, Arendt, and Habermas?" He agrees with Habermas's suggestion: that Heidegger is guilty of "a terrible intellectual hubris" when he suggests that the only proper and authentic response to the supreme danger is to prepare ourselves to watch over unconcealment.

Bernstein next draws attention to an unpublished manuscript of the 1949 lecture that became *The Question Concerning Technology*, which contains the following passage that has been deleted from the published text.

> Agriculture is now motorized food industry – in essence the same as the manufacturing of corpses in gas chambers and extermination camps, the same as blockading and starving of nations [it was the year of the Berlin blockade], the same as the manufacture of hydrogen bombs. (cited in Bernstein 1992: 130)

Bernstein understands this grotesque passage as a natural expression of Heidegger's reaction against the "correct" definition of technology as a neutral instrument which can be used for benign ends of increased food production or the malignant end of extermination of human beings.

> But if we focus on the *essence* of technology then these differences are "non-essential." The manufacturing of corpses in gas chambers more fully reveals the essence of technology . . . Unless we fully acknowledge and confront the essence of technology, even in "manufacturing of corpses in gas chambers," unless we realize that *all* its manifestations are "in essence the same," we will never confront the supreme danger and the possible upsurgence of the saving power. (1992: 131)

Bernstein concludes that the deleted passage is not simply some insensitive remark but rather a necessary consequence of the very way in which Heidegger characterizes *Gestell*, as an unconcealment that claims man and over which he has no control. He sets out a formulaic pattern in Heidgegger's thinking,

> a pattern that turns us away from such "mundane" issues as mass extermination, human misery, life and death, to the "real" plight, the "real" danger – the failure to keep meditative thinking alive . . . It is as if in Heidegger's obsession with man's estrangement from Being, nothing else counts as essential or true except pondering one's *ethos* . . . It becomes clear that the only response that is really important and appropriate is the response to the silent call of Being, not to the silent screams of our fellow human beings . . . when we listen carefully to what he is saying, when we pay attention to the "deepest laws of Heideggerian discourse" then Heidegger's "silence" is resounding, deafening, and damning. (1992: 136)

Bernstein's analysis and conclusions suggests a moral critique of utopian and dystopian theories of internet culture. Although none of the theories I have reviewed is so damned by its inherence arguments as Heidegger's, which blinded him to the *specific* evil of the Holocaust, yet a Postmanian anti-Technopolist may be blinded in parallel fashion to something good or bad about this or that specific aspect of American culture; a Borgmannian may be blinded to something specifically good or bad about some digitally generated artifact; and a gung-ho cybernaut of the Leary persuasion may be blinded to the old-fashioned pleasures of embodiment.

Inherence Instrumentalism

Turkle's category of utilitarian interpretation understands the internet as a tool. The version scouted here under the rubric of inherence instrumentalism interprets the internet as essentially a metatool for creating tools. This general idea derives from Robert Nozick's discussion of a libertarian utopia in *Anarchy, State and Utopia*. Although he did not have the internet in mind, what he says there about a framework for utopia transfers quite naturally to the internet, as well as having greater plausibility there than in political philosophy for the real world.

The internet does not have a culture of simulation, on this metatool account, because it is a tool for creating a *variety* of subcultures, some of which may fit Turkle's description of internet culture, some of which will not, not to mention the variety of internet activity, like setting up a webpage for lecture notes, that does not amount to creating a subculture. The internet is the Swiss Army knife of information technology.

Libertarians sometimes think of *utopia* in this way: ideally, everyone would be free – would have the Lockean "natural right" – to migrate or emigrate as he or she chose. The worlds that result from such to-ing and fro-ing they call *associations.* Acknowledging that there is no single world that's everyone's perfect cup of tea, the libertarian is inspired by a utopia which is a set of possible worlds, with permeable borders, in which one world is the best imaginable for each of us. Those whom you would have in your ideal world are also free to imagine and relocate, perhaps to a world of their own imagining. There could be an incessant churn of relocation, all worlds being ephemeral, or some stable worlds might emerge in which everyone would choose to remain. There will be no one in a stable association who wants out, and no one will be in whose presence is not valued by the others. Libertarianism may be bad politics, but its conception of utopia is a plausible model of the internet.

The claim that inherence instrumentalism makes to being "value free" is provocative, defying a post-Weberian tradition of deconstructing such claims with a view to revealing hidden value commitments, an argumentative strategy that bears Heidegger's imprimatur, as noted above. It may be helpful to clarify the claim with an analogy to a box of paints and a variety of paintings made with them, some of them good paintings, some of them bad, some of them so-so. It would be a logical error, a "category mistake" in Rylean terminology, to evaluate the box of paints as a good, poor, or so-so painting. It is not a painting at all. Classification of the internet as a metatool aims at a similar conclusion. Corresponding to the variety of paintings in the analogy is the variety of *content* on the internet. None of this content is value free in the sense that is being reserved for the internet as a metatool. Content in the middle of the continuum from poor to good might be deemed value free in the sense that it excites no judgments of praise or condemnation with respect to this or that value; such internet content might be described as *bland*. But the sense in which the internet is value free is not like this. Rather, it is like the freedom of the box of paints from being judged a good, bad, or so-so painting. It is not a bland painting, and the internet is not bland internet content, on the inherence instrumentalist account. Inherence dystopians and utopians purport to find something deeply good or bad about the internet, but on an instrumentalist diagnosis either they become so deep that they loose touch with the truth, as illustrated by the attempt to tie the computer inevitably to a society of technocratic administration, or else they are guilty of a part–whole fallacy, judging the whole internet by some of its uses. Even if all

uses had some bad value or effect *X*, that would ground only a balance-of-reasons judgment that one should or should not use the internet, depending on whether *X* outweighs the good value or effect *Y*.

Conclusion

To illustrate once more the dystopian/instrumental/utopian continuum and the balance/inherence vectors that can be traced by reference to it, consider the changes being wrought in work and leisure by the computer revolution. Offices have been transformed by the computer over the past two decades, while web surfing, computer gaming, and internet chatrooms have become significant leisure activities. As recent events in Afghanistan testify, even war, that most regrettably necessary form of work, must be fought with sophisticated information technology in order to achieve success in the battlefield of the twenty-first century; the leisure activity of correspondence is migrating from the pen and the typewriter to computer email, a transition from manipulating matter to manipulating digital bytes that is as significant as any preceding revolution in communication technology. Despite the uneven track record of "dot.coms," business activity on the internet is starting to take giant strides; new communities are being formed on the internet, like Multi-User Dungeons (MUDs), Internet Relay Chat (IRC), and so on, online "third places" between work and home that allow users of the Net a respite from the demands of office and household. Work as traditional as farming is becoming reliant on the boost to organization and efficiency that computers make possible; games like chess, go, poker, and bridge are just as likely to play out on the internet as in physical spaces. Computers and the internet are opening up new employment opportunities, new tools, and new media for artists; correspondingly, creating and maintaining a personal webpage has become an art that many pursue in their free time. Telecommuting and teleconferencing are becoming more widespread, with potentially enormous implications for city design and transportation systems; making friends is no longer channeled by physical neighborhood, and with the development of automatic translation software a great obstacle to cross-cultural friendships, namely lack of a common language, is being removed. New motivations and organizational structures for work are being discovered on the internet, notably the "open source" initiative associated with Linus Torvalds, Eric Raymond, and a legion of true hackers, showing how psychic rewards can replace monetary ones in high-quality software development within the internet milieu; if work is understood as paid employment, contributions to such software development is not work, whereas if it is understood as activity that is instrumental to some further end, such as a new Linux kernel, it is work calling for a high level of skill. This raises the question whether the suffusion of IT into work and leisure will eventually lead to their transcendance in "meaningful work" that is pursued because of its intrinsic motivations, not extrinsic ones such as money. Is there something about information technology that makes it *inherently* amenable to meaningful work? The case could be made that will do so by following a negative and a positive path. The *via negativa* is the elimination of "agonistic work," work that one would gladly avoid if it weren't necessary. The *via positiva* is the creation of attractive environments in which one is always able to work "just as one has a mind." Marx had such an environment in mind when he speculated about the higher stages of communism, in which the division of labor characteristic of capitalism has been overcome and one's distinctively human powers are fully realized, without the compulsion of necessity. In *The German Ideology* he made the point like this:

> For as soon as the distribution of labor comes into being, each man has a particular, exclusive sphere of activity which is forced upon him and from which he cannot escape. He is a hunter, a fisherman, a shepherd, or a critical critic, and must remain so if he does not want to lose his means of livelihood; while in communist society, where nobody has one exclusive sphere of activity but each can become accomplished in any branch he wishes, society regulates the general production and thus makes it possible for me to do one thing today and

> another tomorrow, to hunt in the morning, fish in the afternoon, rear cattle in the evening, criticize after dinner, just as I have a mind, without ever becoming hunter, fisherman, shepherd, or critic. (Tucker, ed., 1974: 124)

Add to Marx's flight of fancy the thought that information technology will be the means by which "society regulates the general production," and you have a form of inherence utopianism about IT. However, given the failure of command economies in real-world tests such as the USSR, Heideggerian inherence dystopianism may recommend itself instead. IT will have taught us, on this account, to view nature as so much "standing reserve" and not even the overcoming of the division of labor will protect us from a mental architecture that we should want to avoid. Another inherence-dystopian option argues that a core value of our civilization, to which our self-respect is inexorably tied, is agonistic work; IT, by showing us how to eliminate such work, will have the unintended consequence of removing the bases of our self-esteem. The aspect of technological determinism is noticeable in these three options. An alternative is the outlook that Karl Popper advocated in *The Open Society and its Enemies* and elsewhere, which views with suspicion ideas about the necessity of history's unfolding and recommends instead that opportunities for change be monitored for unintended consequences, so that choices can be made that reflect knowledge of where change is going wrong. The current debate about genetically modified foods is an example of such monitoring; it also illustrates a tendency for inherence voices to emerge at the dystopian and utopian extremes.

The Popperian outlook may be viewed as contributing to an inherence-instrumentalist interpretation of internet culture, wherein the metatool character of the technology acknowledges dystopian fears and utopian hopes with respect to particular content. At the metalevel, however, the internet is neither good nor bad nor in-between; at the level of specific content, it may be any of these things. The Popperian contribution theorizes the internet, not as historical inevitability to be deplored or valorized holus-bolus, but rather as a locus of possibilities, to be monitored carefully in order to make practically wise choices about its use. As the "Mount Carmel Declaration on Technology and Moral Responsibility" observed in 1974 in its eighth article, "We need *guardian disciplines* to monitor and assess technological innovations, with especial attention to their moral implications" (Hester & Ford 2001: 38). No technology is morally neutral if that means freedom from moral evaluation. But there is no *inherent* reason why that evaluation should be pro or con.

Acknowledgments

I wish to thank Luciano Floridi for invaluable philosophical and editorial comments about earlier drafts of this chapter.

References

Bernstein, Richard. 1992. The New Constellation: The Ethico-political Horizons of Modernity/Postmodernity. Cambridge, MA: MIT Press. [For graduates; a lucid exposition and critique of Heidegger's philosophy of technology.]

Birkerts, Sven. 1994. *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. London: Faber & Faber. [For the educated public. It evokes the depth and inwardness that reading can obtain, and expresses concern about the future of reading in an information age.]

Borgmann, Albert. 1999. *Holding on to Reality*. Chicago: University of Chicago Press. [For graduates and ambitious undergraduates. He makes an engaging case for focal values in life that keep us close to nature and our communities.]

Deutsch, David. 1997. *The Fabric of Reality*. Harmondsworth: Penguin. [For graduates and ambitious undergraduates. He brings virtual reality to the center of his interpretation of quantum mechanics.]

Evans, Fred. 1993. *Psychology and Nihilism: A Genealogical Critique of the Computational Model of Mind*. New York: SUNY Press. [For graduates. This is an interpretation of cognitive psychology inspired by Nietzsche.]

Heidegger, Martin. 1973. "Overcoming metaphysics." In *The End of Philosophy*. New York: Harper and Row. [For graduates. Bernstein draws

on this essay for his interpretation of Heidegger's philosophy of technology.]

——. 1977. "Letter on humanism." In *Martin Heidegger, Basic Writings*. New York: Harper and Row. [For graduates. Bernstein draws on this essay for his interpretation of Heidegger's philosophy of technology.]

——. 1977. *The Question Concerning Technology*. New York: Harper and Row. [For graduates. This is the inspiration for many interpretations of technology in recent decades.]

Heim, Michael. 1993. *The Metaphysics of Virtual Reality*. Oxford: Oxford University Press. [For undergraduates. He writes vividly about how we are being transformed by information technology, and about the long-term prospects for virtual reality.]

Hester, D. Micah and Ford, Paul J. 2001. *Computers and Ethics in the Cyberage*. Englewood, NJ: Prentice-Hall. [For undergraduates. This is typical of many accessible anthologies of essays on technology and computers from authors in a variety of fields.]

Jameson, Fredric. 1991. *Postmodernism, or, The Cultural Logic of Late Capitalism*. Durham, NC: Duke University Press. [For graduates. This is an important source of theory about postmodernism.]

Kelly, Kevin. 1994. *Out Of Control: The New Biology of Machines, Social Systems, and the Economic World*. New York: Addison-Wesley. [For undergraduates and the educated public. This is imaginative, mind-stretching scientific journalism about convergence of the natural and the human-made.]

Lovelock, James. 1988. *The Ages of Gaia: A Biography of our Living Earth*. New York: W. W. Norton. [For undergraduates. Kelly (see above) weaves Lovelock's ideas into his vision of where the Earth is heading.]

Mazlish, Bruce. 1993. *The Fourth Discontinuity: The Co-evolution of Humans and Machines*. New Haven and London: Yale University Press. [For graduates. This is an historian's scholarly account of the convergence between humans and machines.]

Nozick, Robert. 1974. *Anarchy, State and Utopia*. New York: Basic Books. [For graduates. This is a remarkable defense of libertarianism that poses hard questions for liberals and Marxists, while painting a minimal state as a veritable utopia.]

Popper, Karl. 1971. *The Open Society and its Enemies*. Princeton: Princeton University Press. [For the educated public. This is his contribution to the war effort in the Second World War, tracing communism and fascism to philosophical roots in Plato, Hegel, and Marx.]

Postman, Neil. 1993. *Technopoly: The Surrender of Culture to Technology*. New York: Knopf. [For the educated public. This is an articulate call to arms against the influence of technology on culture.]

Slouka, Mark. 1995. *War of the Worlds*. New York: Basic Books. [For the educated public. He argues that computer games like MUDs are making us lose touch with reality.]

Tipler, Frank J. 1994. *The Physics of Immortality*. New York: Doubleday. [For graduates. Deutsch (see above) draws on Tipler for some of his cosmological speculations.]

Tucker, Robert C. 1972. *The Marx–Engels Reader*. New York: W. W. Norton & Co. [For graduates. This is typical of many anthologies of Marx's and Engels' writings.]

Turkle, Sherry. 1995. *Life on the Screen*. New York: Simon & Schuster. [For the educated public. This is an indispensable interpretation of the cultures that have grown up around the computer.]

Woolley, Benjamin. 1992. *Virtual Worlds*. Oxford: Blackwell. [For the educated public. This is a witty and philosophically informed series of essays on matters pertaining to virtual reality.]

Chapter 8

Digital Art

Dominic McIver Lopes

Introduction

Artworks are artifacts, their making always involves some technology, and much new art exploits and explores new technologies. There would be no novels without inexpensive printing and book binding. The modern skyscraper is a product of steel manufacture. Jazz married the European technology of the diatonic scale to African rhythms. A factor in the origins of Impressionism was the manufacture of ready-made oil paints in tubes, which facilitated painting outdoors in natural light. As soon as computers became available, they were used to make art – the first computer-based artwork was created as early as 1951 (Reffen Smith 1997: 99) – and since then the body of digital artworks has grown by leaps and bounds. But although the first philosophical paper on "cybernetic art" appeared in 1961 (Parkinson 1961), philosophers are only now beginning to address in depth the questions raised by digital art. What is digital art? How, if at all, is it new and interesting as an art medium? Can it teach us anything about art as a whole?

Answering these questions provides an antidote to the hype that frequently attaches to digital art. We hear that computer art is overhauling our culture and revolutionizing the way we think about art. It frees artists from the materiality of traditional art media and practices. Art appreciators, once passive receptacles of aesthetic delight, may finally participate actively in the art process. Pronouncements such as these spring less from careful study and more from marketing forces and simple misunderstandings of a complex and multifaceted technology. An accurate conception of the nature of digital art and its potential may channel without dousing the enthusiasm that attends any innovation. At the same time, it counterbalances some cultural critics' jeremiads against digital art. Radical antihype often depends for its rhetorical force on our reaction to hype. When we are told that electronic music or fractal art or virtual-reality goggles are *the* future of art, we are given good reason to doubt the credibility of our informant and this doubt may engender blanket skepticism about digital art. But while most digital art is admittedly dreadful, this does not show that it never has value or interest. The correct lesson to draw is that we should proceed with caution.

This chapter is divided into three sections. The first reports on the use of computers as tools in art-making. The second describes some artworks that capitalize on the distinctive capabilities of digital computers and digital networks. To make sense of these works we must define digital art and consider whether it is a new art medium. The third reviews the use of computers as instruments that yield general insights into art-making. This

three-section division is one case of a useful way of thinking about any use of computers, not just in the arts. For example, a philosophy of artificial intelligence might begin by discussing computers as cognitive aids (e.g. to help with calculations), then consider whether computers possess a kind of intelligence, and close with a discussion of the use of computer models of the human mind in cognitive psychology.

1 Making Art Digitally

The digital computer has occasioned two quite distinct kinds of innovation. It has automated and sped up many tasks, especially routine ones, that were once relatively difficult or slow. It has also made some activities possible that were previously impossible or else prohibitively difficult. Most discussions of digital art are captivated by the latter kind of innovation; however, the impact of the former should not be ignored. If art always involves some craft then the practice of that craft may incorporate the use of computers. Moreover, a clear view of the uses of computers as art-making tools can help crystallize a conception of the kind of innovation that involves opening up new possibilities for art.

When the craft underlying an art medium has practical, non-art applications, digital technology is frequently brought to bear to make the exercise of that craft easier and more efficient. Here the use of computers in making art simply extends their use in other areas of human endeavor. The first computer imagining technologies, output plotter drawings, were developed for engineering and scientific uses, but were quickly adopted by artists in the early 1960s. It hardly needs to be pointed out that word-processors have proved as much a boon to literary authors as to office managers. Software created for aeronautical design paved the way for the stunning, complex curves that characterize Frank Gehry's recent buildings, notably the Guggenheim Bilbao. Since digital sound processing and the MIDI protocol were developed specifically with music in mind, music is an exception to the rule that digital art technologies adapt technologies fashioned for some non-art purpose. In each of these cases, however, the computer merely realizes efficiencies in art-making or art distribution. Digital technology, including digital networking and the compact disk, is used to store music, as did vinyl records, but in a format that is considerably more portable and transmissible without introducing noise. Musical recordings that once required live musicians, a studio, and several technicians, can now be made at a fraction of the cost by one person in her garage with a keyboard and a computer.

Computers sometimes make it easier for artists to work and, by reducing the technical demands of the craft underlying an art medium, they sometimes make it easier for untutored novices to make art. In addition, some uses of computers in making and distributing art cause artworks to have properties they would not otherwise have. The use of typewriters by some modernist writers in the early twentieth century influenced the character of their writing. Relatively inexpensive digital movie editing encourages film-makers to experiment with faster pacing and more complex sequencing. Poor musical technique is now no barrier to recording music and distributing it worldwide from one's desk. Tod Machover's hyperinstruments can be played in interesting ways – some, for instance, are soft toys whose sound depends on how they are squeezed – and can be used to make music whose sound reflects its instrumentation (see <http://www.media.mit.edu/hyperins>). What properties artworks of an era possess depends in part upon the technologies employed in making art during that era. Art's history is partly driven by technological innovation.

While the kinds of innovations discussed so far generate artworks with new properties, they neither beget new art media nor change our standards for evaluating artworks. An aesthetic evaluation of a performance of a pop song need not take into account whether the recording of it is analog, digitally remastered, or direct-to-digital, and whether that recording is played back from a vinyl record, a reel of magnetic tape, a compact disk, or an MP3 file. The relative ease of online publication means that much more is published, but the nature of literature and its aesthetically relevant properties endure. A novel is a novel and is as good or as bad as it is whether

it is printed and bound into a book or emailed to one's friends. It is important to recognize how computers have found their way into artists' studios – or made the resources of a studio more widely and cheaply available. But this is no revolution in the nature of the arts.

2 The Digital Palette

Computers ease the performance of some tasks but they also equip us to undertake new tasks. Exploiting this, artists may invent new varieties of art, including what we may designate the "digital arts." One question to be answered is what is characteristic of digital art media. Theorists typically propose that digital art is novel in two ways, the first deriving from virtual-reality technologies and the second deriving from the capacity of computers to support interactivity. Something must be said about what virtual reality and interactivity are, and it will be helpful to describe some artistic uses of each. But since our goal is to devise a theory of digital art, it is prudent to begin by considering what an adequate theory of any art medium should look like.

Art media are species of a genus that comprises all and only works of art. This genus can be characterized either evaluatively or descriptively. According to evaluative characterizations, works of art are necessarily good as works of art, and "art" is an essentially honorific term. Some theorists who write about digital art (especially its critics) have this characterization in mind. Brian Reffen Smith, himself a computer artist, dismisses much of what goes under the banner of digital art as "graphic design looking a bit like art" (Reffen Smith 1997: 102). He does not allow that the works in question are poor art, for art, he assumes, is necessarily good as art. Descriptive conceptions of art allow that some works may be failures as works art and yet deserve the name, so that to call something "art" is not necessarily to commend it but merely to acknowledge its membership in the class of artworks, good and bad. It is a matter of considerable controversy how to characterize the conditions of membership in this class (see Carroll 2000, Davies 2000). Fortunately, consensus is not necessary if our aim is to characterize digital art. We may assume that digital art is a kind of art and concentrate our efforts on what distinguishes it from other kinds of art. And although we may proceed with either an evaluative or descriptive characterization of art, it is wiser to characterize digital art as art in the descriptive sense, so as not to beg any questions about its quality.

The assumption that digital art should be considered art, even when art is characterized descriptively, is not uncontroversial. One theorist asks of digital graphic art,

> whether we should call it "art" at all. In treating it as art we have tended to weigh it down with the burden of conventional art history and art criticism. Even now – and knowing that the use of computing will give rise to developments that are as far from conventional art as computers are from the abacus – is it not too late for us to think of "computer art" as something different from "art"? As something that perhaps carries with it parallel aesthetic and emotional charges but having different and more appropriate aims, purposes and cultural baggage? (Lansdown 1997: 19)

There are two reasons that this objection should not give us pause, however. Even granting that what we count as art depends on a welter of social practices and institutions, art status is not a matter for deliberate legislation. More importantly, the objection misses an important fact about art. We never judge or see an artwork merely as art but always as some kind of artwork – as belonging to some art medium. If digital art is art, it remains an open question whether it is an art medium that inherits the history, purposes, standards of criticism, and "cultural baggage" of any other art media.

In his classic paper "Categories of Art," Kendall Walton maintains that we perceive every work of art as belonging to some category of art, where art categories are defined by three kinds of properties: standard, variable, and contrastandard properties (Walton 1970). Standard properties of works in a category are ones in virtue of which they belong to the category; lacking a feature standard for a category would tend

to disqualify a work from the category ("having an unhappy ending" is a standard property of tragedies). We discriminate among works in a category with respect to their variable properties ("featuring an indecisive prince" is a variable property of works in the category of tragedies). Contrastandard properties of works with respect to a category are the absence of standard features in respect of the category. A tragedy may have the contrastandard feature of having an ending that is not unhappy. But why perceive a work in a category when it has properties that are contrastandard with respect to that category? For Walton, at least four factors determine what category we should perceive a work as belonging to: the work's having a relatively large number of properties that are standard for the category, the artist's intention or expectation that the work be perceived as in the category, the existence of social practices that place it in the category, and the aesthetic benefits to be gleaned from perceiving the work as being in the category – a drama with a happy ending that is inventive, even shocking, when viewed as tragedy, may seem old hat when viewed as comedy.

Art categories provide a context within which we appropriately interpret and evaluate artworks. To appreciate a work of art one must know how it resembles and differs from other works of art, but not every resemblance or difference is aesthetically significant. There is only a point to noticing differences among works that belong to a kind and to noticing similarities among works when the similarity is not shared by everything of its kind. Acid jazz differs from opera, but to appreciate John Scofield's "Green Tea" as a work of acid jazz it is not enough to hear how it differs from *Rigoletto* – one must recognize how it differs from other works of acid jazz.

Moreover, what properties are standard, contrastandard, and variable with respect to a category is subject to change. Suppose that it is a standard property of photography that photographs accurately record visible events. As the use of software for editing rasterized photographs increases, this may become a variable property of the category. Digital image doctoring may thereby change how we see all photographs (Mitchell 1992; Savedoff 1997). The lesson is that contexts within which we appreciate and evaluate works of art are fluid and can be shaped by technological forces.

As the examples given indicate, there are several schemes of categories into which artworks can be portioned. One scheme comprises the art media – music, painting, literature, theater, and the like. Another scheme comprises genres of art, such as tragedy and melodrama; works in these categories may belong to different art media. A third scheme, that of styles, also cuts across media and genres. There are postmodernist parodies and postmodernist comedies; some of the former are musical while others are architectural and some of the latter are literary while others are pictorial. How, then, should we characterize the scheme of art categories that comprises the art media? For it is within this scheme that we might expect to make room for a category of digital art.

One way to characterize the art media is with reference to their physical bases. Musical works are sounds; pictures are flat, colored surfaces; and theatrical performances consist in human bodies, their gestures and speech, together with the spaces in which they are located. Indeed, we use the term "medium" ambiguously to name an art form and its physical embodiment. The "medium of pictures" can denote the pictorial art form or it can denote the stuff of which pictures are made – oil paint, acrylic, encaustic, ink, and the like. Nevertheless, ordinary usage notwithstanding, we should distinguish art media from what I shall call, following Jerrold Levinson, their "physical dimensions" (Levinson 1990: 29). The reason is that works in different art media may share the same physical dimension and works in the same art medium may have different physical dimensions. The case of literature is instructive. Literary works can have many physical dimensions, for they can be recited from memory as well as printed on paper. Moreover, when novels are printed on paper they have the same physical dimension as many pictures, but although some artworks are both literary and pictorial (visual poems for instance), printed volumes of *Lady Chatterley's Lover* are not pictures.

The medium of literature is independent of any particular physical dimension because works of literature are made up of bits of language and language is independent of any particular physical

dimension. Yet there is a sense, however stretched, in which every art medium comprises a "language," understood as embodied in a set of practices that govern how the materials of the medium are worked. This is all we need in order to characterize the art media. Artworks standardly belong to the same art medium when and only when they are produced in accordance with a set of practices for working with some materials, whether physical, as in sculpture, or symbolic, as in literature. These materials together with the practices of shaping them determine what works are possible in an art medium. Call the materials and the practices governing how they can be worked the art medium's "palette."

The digital palette comprises a suite of technologies and ways of using them that determine what properties digital artworks can possess, including those properties that are standard and variable with respect to the category of digital art. Since computers can be programmed to serve indefinitely many tasks, the digital palette is unbounded. But we can discern, if only in outline, some of the potential of the digital palette by canvassing some typical cases of innovative digital art. We should keep in mind throughout that the point of thinking of digital artworks as belonging to a digital art medium is that we properly appreciate and evaluate digital artworks only when we perceive them within the category or medium of digital art, as it is characterized by the digital palette.

One digital technology that is much discussed in recent years among media theorists and that is thought to engender a new digital art form is virtual reality. This is standardly defined as a "synthetic technology combining three-dimensional video, audio, and other sensory components to achieve a sense of immersion in an interactive, computer-generated environment" (Heim 1998: 442). The vagueness of this definition accurately reflects the wide range of technologies that are called virtual reality. "Three-dimensional video" can denote the use of perspective animations to represent three-dimensional scenes on two-dimensional computer monitors, often with exaggerated foreshortening (as in most computer games), or it can denote the use of stereoscopic animations viewed through virtual-reality goggles. The question to ask is whether virtual reality makes possible an art medium with distinctive properties.

Some claim that virtual reality uniquely generates an illusion that the user is in the computer-generated environment, perceiving it. But what is meant by "illusion"? On the one hand, it does not appear that even the most sophisticated virtual-reality set-ups normally cause their users to believe, mistakenly, that they are part of and perceiving the computer-generated environment. On the other hand, any imagistic representation elicits an experience like that of perceiving the represented scene, even images (e.g. outline drawings) that are far from realistic. Virtual reality could be redescribed without loss as "realistic imaging" and classified with other realistic imaging such as cinema or three-imensional (stereoscopic) cinema. If virtual reality offers anything new it is the possibility for interaction with the occupants and furniture of the computer-generated environment. As Derek Stanovsky puts the point, "computer representations are different because people are able to interact with them in ways that resemble their interaction with the genuine articles" (see Chapter 12, VIRTUAL REALITY). Virtual reality as realistic imaging should not be confused with interactivity.

The interactivity of computers capitalizes on their ability to implement complex control structures and algorithms that allow outputs to be fine tuned in response to different histories of inputs. What properties a work of interactive digital art possesses depends on the actions of its user. The point is not that every user has a different experience when engaging with an artwork – that is arguably true of our experiences of all artworks. The point is rather that the structural properties of the work itself, not just how our experience represents the work, depend on how we interact with it (Lopes 2001). Defined in this way, digital interactive art is something new and it exists precisely because of the special capabilities of computing technology.

A hypertext story, such as Michael Joyce's widely read *Afternoon, A Story* of 1987, is interactive because it allows the reader to follow multiple narrative pathways, so that the story goes differently on each reading. But there is no reason that hypertext need involve hyperlinked text that the user selects. Simon Bigg's *Great*

Wall of China of 1999 (at <http://hosted.simonbiggs.easynet.co.uk>) transforms a display of the text of Kafka's story in accordance with movements of the user's mouse. The reader of Jeffrey Shaw's 1989 *Legible City* sits on a fixed bicycle which he or she uses to navigate a landscape built of words, each route through the landscape telling a story about a city. Indeed, the input of users to interactive artworks can take a variety of forms: gesture, movement, sound, drawing, writing, and mere physical presence have all been used. Nor is interactive art always narrative in form. Avatar technologies and synchronous remote puppeteering enable users to act in represented performance spaces. Peter Gabriel's *Xplora 1* CD-ROM of 1993 allows its owner to remix Gabriel's music so that it has different sound properties from one occasion of interaction to the next. Robert Rowe's *Cypher* and George Lewis's *Voyager* are computer programs that improvise music in real time as part of an ensemble that includes human musicians. Since what music the computer makes depends on what the other players in the ensemble do, the computer is as interactive as musicians jamming with each other.

One way to see what is special about the works just described is to consider their ontology. Artworks can have, broadly speaking, one of two ontologies. Some artworks, paradigmatically paintings, have a unitary ontology: the work just is the painting, a spatio-temporally bounded particular. Multiple-instance artworks, paradigmatically works of music and literature, have a dual ontology: they are types whose instances are tokens. Most musical performances, for example, are tokens of types that are musical works. The work type determines the properties which anything must possess in order to count as instances of it, yet we apprehend the work through its instances. In the case of music, we typically abstract the musical work from performances of it by stripping from them properties of the performances themselves. This explains how it is possible for a work and its instances to have different as well as shared properties, especially different aesthetic properties. We evaluate performances as aesthetic objects in their own right and yet we evaluate a work performed without thereby evaluating any performance of it. A good work can be given poor performances and a poor work given performances that are, *qua* performances, good but not redeeming.

According to Timothy Binkley, the aesthetically relevant features of a predigital artwork are features of its physical embodiment (Binkley 1997, 1998a, 1998b). To make an artwork is traditionally to "maculate" some physical substance, shaping it into the work. But digital artworks are not physical objects, for the computer "computes abstract numbers with mathematical algorithms rather than plying physical material with manual implements" (Binkley 1998a: 413). Instead of making things, digital artists manipulate data structures; they "mensurate" symbols instead of "maculating" physical stuff. Of course, Binkley realizes that the data structures making up digital artworks always take some physical, usually electronic, embodiment; his point is that the data and its structure is independent of any particular physical embodiment. For this reason digital art "bears no telltale traces of the magnetism, electricity, or cardboard that might happen to host its abstract symbols" (Binkley 1998b: 48). Digital artworks are therefore types. Their aesthetically relevant features are not features of physical objects. They are indefinitely reusable and can be copied with perfect accuracy (think of a digital image sent by email from one person to many others). Binkley concludes that digital art diminishes the importance of art's physical dimension (Binkley 1997: 114; Binkley 1998b: 50). It is, he writes, "an art form dedicated to process rather than product" (Binkley 1998a: 413).

The claim that digital artworks are types is instructive, as is the observation that they are for this reason indefinitely reusable and perfectly reproducible. Also instructive, however, are two related mistakes in Binkley's account. Binkley's first mistake is to take painting's ontology as paradigmatic of all art – that is, by assuming that all nondigital artworks are physical objects. Literature, as we have seen, is a clear counterexample. Musical works, if they are types tokened in individual performances or playings, are another counterexample. When I listen to a performance of "Summertime" I am hearing two things. One is the performance, which is a physical event, and the second is the song itself, which is not identical to the performance though I apprehend its

features by listening to the performance. The case of music indicates Binkley's second mistake. From the fact that digital artwork types are nonphysical it does not follow that their tokens are not physical. Performances of "Summertime" are physical events and our aesthetic interest in them is partly an interest in their physical qualities. Binkley thinks of a computer as a central processing unit and a digital artwork as the data structure a CPU processes. But this ignores two additional and essential components of the computer, the input and output transducers. A digital image is a data structure but it is tokened only by being displayed on an appropriate device, usually a printer or monitor. Indeed, our aesthetic interest in the image is an interest above all in properties of the physical embodiment of its tokens.

David Saltz identifies three design elements essential to digital interactivity: a sensing device (such as a keyboard or mouse) that transduces user actions as inputs, a computational process that systematically relates inputs to outputs, and a display mechanism that transduces outputs into something humanly perceptible (Saltz 1997: 118). All three elements must be in place in order for an interactive piece to vary in its content or appearance with human interaction. For this reason, Saltz models the ontology of interactive art on that of performance art. An interaction is performative, according to Saltz, "when the interaction itself becomes an aesthetic object . . . interactions are performative to the extent that they are *about* their own interactions" (Saltz 1997: 123). The aesthetically relevant properties of performative interactions are properties of the interactor in the work, who plays a role in the interaction's unfolding. But there is no work type of which individual interactions are tokens since the interactions are unscripted, and in the performing arts it is the script (or score or choreography) that identifies individual performances as tokens of one work type. Saltz infers that "to interact with a work of computer art does not produce a token of the work the way performing a dramatic or musical work does" (Saltz 1997: 123).

Neither Binkley's nor Saltz's view adequately describes the ontology of interactive digital art. According to Binkley, only digital work types are objects of aesthetic attention; according to Saltz interactive works are not tokens of aesthetically interesting types. However, the virtue of the application of the type–token distinction to art is that it allows for dual objects of aesthetic attention. We usually attend simultaneously to properties of a performance *qua* performance and to properties of the work performed. The fact that we direct our attention upon interactive processes, or upon our own actions as interactors, does not show that we cannot and do not simultaneously attend to properties of a work type with which we are interacting. Saltz is right that there is no interactive work type understood as what is indicated by a script or score. But it does not follow that we cannot descry features of an interactive work type through instances of interaction with it. The contours of the work type are drawn by what interactions it makes possible. *Afternoon* is many stories, but it is important to know what set of stories it tells and how: these give access to properties of *Afternoon* itself, not the individual stories our interactions with it generate. Moreover, we miss something important if we do not view interaction instances as instances of a work type, since to fully appreciate an interaction as an interaction, one must regard it as means of discerning the work's properties. As one commentator puts the point, "the interactive art experience is one that blends together two individualized narratives. The first is the story of mastering the interface and the second is about uncovering the content that the artist brings to the work" (Holmes 2001: 90).

Interactive work instances are not tokened by performance or playing (as in live and recorded music) and they are not tokened by recital or printing (as in literature); they are tokened by our interaction with them. The way instances of an interactive work are tokened cannot be modeled on the way musical or literary works are tokened. In place of the score, the script, and the text we have the individual user's interaction (Lopes 2001). This is one way of seeing what is new about interactive digital art. It gives a role to its user, not just in interpreting and experiencing the work but in generating instances of it, that users of no other art media enjoy. An interactor tokens an interactive artwork in a way that a reader or spectator of a non-interactive artwork does not.

Interactivity, unlike virtual reality, is distinctive of the digital palette, but not all digital art is interactive. There are many rather more mundane functions that computers perform and that provide resources for the digital palette. Word-processors routinely check the spelling of documents: Brian Reffen Smith has created artworks by first running a text in English through a French spell checker, which substitutes orthographically similar French words for the English originals, and then translating the French words back into English (Reffen Smith 1997: 101–2). So-called interface artworks are applications that change the way familiar graphical user interfaces work. I/O/D's Web Stalker provides an alternative, exploded perspective on websites, for instance (see <http://bak.spc.org/iod>). Like much art of the past century that takes as one of its main subjects the technical basis of its own medium, some digital art uses digital technologies in order to represent or draw our attention to features of the digital art medium.

Early explorations of a new medium tend to imitate the media from which it sprang. Photography aspired at first to the look of painting and it was only after several decades that photographers made unabashedly photographic photographs. One explanation for this is that a new medium must establish its status as art by associating itself with a recognized art medium. Another explanation is that it is difficult to discern the full potential of a medium's palette in advance of actually using it to make art. Whatever the explanation, it is only with time that we can expect digital art to look less like other kinds of art and to acquire a character of its own. This process involves coming to see what standard and variable properties characterize the digital medium and how they are determined by the digital palette. It culminates in our evaluating digital art on its own terms, as digital art.

3 Computing Creativity

Making art is a cognitive activity, as well as a physical and a social activity. Just as philosophers and behavioral scientists study cognitive processes such as vision or language acquisition by developing computer models of those processes, they may learn about the cognitive underpinnings of art-making by building art-making computers. Computers have been programmed as a means to learn about drawing, musical composition, poetic writing, architectural style, and artistic creativity in general.

One may immediately object to the viability of this enterprise. Artworks are necessarily artifacts and artifacts are the products of intentional action, but if "art"-making computers have no intentions, then they cannot make artworks. If they cannot make artworks, it is pointless to use them to study art-making processes. The objection does not assume that no computers or robots can have intentions. It assumes only that the computers that have been programmed to make "art" are not intentional agents – and this is a plausible assumption. The drawing system described below can be downloaded from the internet and installed on a computer that can otherwise do nothing more than send email and word-process.

Granting that artworks are intentionally made artifacts, two replies can be made to this objection. The first challenges the objection directly by arguing that computer-made "art" is art indeed. Typical acts of art-making involve two intentions: an artist intends to make an object that has certain intrinsic properties (e.g. a given arrangement of colors, a meaning) and further intends, typically through the realization of the first intention, to make a work of art. Distinguishing these intentions makes sense of some atypical acts of art-making. An artist selects a piece of driftwood, mounts it, and labels it (alluding to Duchamp's snow shovel) *Notes in Advance of a Broken Arm*. If *Notes* is a work of art, it is a work of art in the absence of an intention to create an object with the physical features possessed by the driftwood. *Qua* driftwood, the object is not an artifact, yet it is an artifact *qua* artwork, since it is mounted and displayed with the intention that it be a work of art. We may view a drawing made by a computer as, like the driftwood, shaped by a force of nature, and yet deem it art since we intend that it be displayed as art. The second reply concedes that computer-made "art" is not art but suggests that it is quasi-art instead. Computer drawing is sufficiently like human drawing that we can use

the former to study the latter. We cannot use what a computer does to study that part of the art-making process that depends on agency or on social institutions, but that is no limitation we need worry about.

Early experiments in computer creativity extend a venerable tradition of automatic art. Wind chimes or aeolian harps are designed to make music, but the particular music they make is not composed. Humans can be involved in making automatic art when they do nothing more than implement an algorithm. Mozart's "Musikalisches Würfelspiel" requires its players to role dice that determine how the music goes. In the surrealists' game of Exquisite Corpse, each player draws on part of a surface the rest of which is blocked from view, making part of an image that, as a single image, nobody drew. During the 1950s and 1960s, the heyday of "systems art," composers such as Iannis Xénakis and John Cage created algorithms for music generation that were implemented on computers. A currently popular form of automatic art is genetic art, in which a computer randomly propagates several mutations of a form, of which humans select one, that provides the material for another round of mutation and selection (e.g. Sims 1991).

Clearly, not all computer-based automatic art illuminates processes of human art-making. What is required is, first, that the computer's computational architecture be designed to model that of humans, at least at relatively high levels of abstraction, and, second, that the choice of algorithms be constrained so as to produce works that resemble those made by humans. Whereas automatic art looks, sounds, or reads like automatic art, art made by computers designed to model human art-making should pass an aesthetic version of Turing's imitation game.

Harold Cohen's AARON, a version of which can be installed as a screen saver on personal computers, draws convincing figures – figures that are sufficiently charming that they have been exhibited in art galleries (see <http://www.kurzweilcyberart.com>). The system's four-component architecture reflects some of what we would have to know in order to understand how we make images (Burton 1997). AARON possesses a way of creating physical images, either by coloring pixels on a screen or by sending data to a printer or a plotter. It has also been supplied with a set of "cognitive primitives" – the basic elements of line pattern and coloration that form the universal building blocks of pictures. A set of behavioral rules governs how the system deploys the cognitive primitives in response to feedback from the work in progress. Finally, a second set of behavioral rules directs the system's work in light of knowledge of how things look in the world – knowledge, for instance, of human anatomy. While these rules might be devised so as to produce only realistic images in canonical perspective, AARON is able to produce images that fit into a variety of human drawing systems, including those favored by children of different ages.

AARON models an isolated artist, one who works outside a drawing tradition. David Cope's EMI is designed to write music that mimics music in the style of historical composers on the basis of "listening" to a selection of their work (Cope 1991). EMI's top-level algorithm comprises six steps: encoding input works by a target composer into a format it can manipulate, running a pattern matcher on the input, finding the patterns that make up the composer's stylistic "signature," composing some music in accordance with an appropriate set of rules, overlaying the composer's "signature" upon the newly composed music, and finally adding musical textures that conform to the composer's style. The technologies employed include rule-based expert systems, pattern recognition neural nets, LISP transition networks, and a style dictionary. The results are remarkable: expert audiences are unable to reliably distinguish EMI's versions of music in the styles of Mozart and Rachmaninoff from the originals.

Specialized applications of this technology enable systems to improvise music in real time with or without human musicians. These systems incorporate real-time listening, musical analysis, and classification with real-time music generation. Moreover, since the music generated at a given time must be recognizably related in an appropriate style to earlier elements of the piece, these systems have been developed in tandem with computational theories of improvisation (Johnson-Laird 1993). Analogous style recognition and art-production systems have been designed for architecture (e.g. Stiny & Mitchell 1980) and poetry. Here is a haiku written by

Ray Kurzweil's Cybernetic Poet in imitation of the style of Wendy Dennis (<http://www.kurzweilcyberart.com/poetry/rkcp_poetry_samples.php3>):

Page

Sashay down the page
through the lioness
nestled in my soul

Supposing AARON, EMI, and the Cybernetic Poet make art, or quasi-art, it does not follow that their activities are creative. This means it is possible to study what creativity is by considering the possibility of creative computers. Margaret Boden approaches the topic of creativity in science and art by asking: can computation help us understand creativity? can computers appear creative? can computers appear to recognize creativity? can computers be creative? (Boden 1994: 85; Boden 1998). The point is not to answer these questions primarily so as to understand the capabilities of computers but rather so as to gain a deeper understanding of creativity itself.

Boden, for example, draws a distinction between historical creativity, a property of a valuable idea that nobody has ever had before, and psychological creativity, a property of a valuable idea that could not have arisen before in the mind of the thinker who has the idea (Boden 1994: 76). Computers can clearly originate historically creative ideas; it is their capacity for originating psychologically creative ideas that is in question. To resolve this question we need to know what it means to say an idea "could not" have arisen before in a thinker. A creative idea is not merely a novel idea in the sense that a computational system is said to be able to generate novel outputs. I have never before written the previous sentence but the sentence is hardly creative, for my capacity to write the sentence is a computational capacity to generate novel sentences. Boden proposes that a system is creative only when it can change itself so as to expand the space of novel ideas it is capable of generating. In order to change itself in this way, it must represent its own lower-level processes for generating ideas and it must have some way of tweaking these processes. Genetic algorithms, which enable a system to rewrite its own code, appear to meet these conditions, and so suggest one way in which computers can be made to be genuinely creative. What is important here is not the ultimate adequacy of Boden's account but its value as an illustration of the prospects of developing a theory of creativity by modeling it computationally.

It is tempting to assume that the cutting-edge applications of digital technologies are exclusively scientific or industrial. Artists have explored the potential of computers since their invention, sometimes using them in surprising ways. We might learn something about computers from their use by artists. Yet a great deal of computer-based art is pure techno-spectacle that has not much more to offer us than the shiny newness of its technology. Digital technology is as much a challenge as well as an opportunity.

References

Binkley, T. 1997. "The vitality of digital creation." *Journal of Aesthetics and Art Criticism* 55(2): 107–16.

——. 1998a. "Computer art." In M. Kelly, ed., *Encyclopedia of Aesthetics*, vol. 1. Oxford: Oxford University Press, pp. 412–14.

——. 1998b. "Digital media: an overview." In M. Kelly, ed., *Encyclopedia of Aesthetics*, vol. 2. Oxford: Oxford University Press, pp. 47–50.

Boden, M. 1994. "What is creativity?" In M. Boden, ed., *Dimensions of Creativity*. Cambridge, MA: MIT Press.

——. 1998. "Computing and creativity." In T. W. Bynum and J. H. Moor, eds., *The Digital Phoenix: How Computers are Changing Philosophy*. Oxford: Blackwell.

Burton, E. 1997. "Representing representation: artificial intelligence and drawing." In S. Mealing, ed., *Computers and Art*. Exeter: Intellect.

Carroll, N., ed. 2000. *Theories of Art Today*. Madison: University of Wisconsin Press.

Cope, D. 1991. *Computers and Musical Style*. Madison: A-R Editions.

Davies, S. 2000. "Definitions of art." In B. Gaut and D. M. M. Lopes, eds., *Routledge Companion to Aesthetics*. London: Routledge.

Heim, M. 1998. "Virtual reality." In M. Kelly, ed., *Encyclopedia of Aesthetics*, vol. 4. Oxford: Oxford University Press, pp. 442–4.

Holmes, T. 2001. "Rendering the viewer conscious: interactivity and dynamic seeing." In R. Ascott, ed., *Art, Technology, Consciousness, mind@large*. Exeter: Intellect.

Johnson-Laird, P. 1993. "Jazz improvisation: a theory at the computational level." In P. Howell, R. West, and I. Cross, eds., *Representing Musical Structure*. London: Academic.

Lansdown, J. 1997. "Some trends in computer graphic art." In S. Mealing, ed., *Computers and Art*. Exeter: Intellect.

Levinson, J. 1990. "Hybrid art forms." In *Music, Art, and Metaphysics*. Ithaca: Cornell University Press.

Lopes, D. M. M. 2001. "The ontology of interactive art." *Journal of Aesthetic Education* 35(4): 65–82.

Mitchell, W. J. 1992. *The Reconfigured Eye: Visual Truth in the Post-Photographic Era*. Cambridge, MA: MIT Press.

Parkinson, G. H. R. 1961. "The cybernetic approach to aesthetics." *Philosophy* 36: 49–61.

Reffen Smith, B. 1997. "Post-modem art, or: virtual reality as trojan donkey, or: horsetail tartan literature groin art." In S. Mealing, ed., *Computers and Art*. Exeter: Intellect.

Saltz, D. Z. 1997. "The art of interaction: interactivity, performativity, and computers." *Journal of Aesthetics and Art Criticism* 55: 117–27.

Savedoff, B. E. 1997. "Escaping reality: digital imagery and the resources of photography." *Journal of Aesthetics and Art Criticism* 55: 201–14.

Sims, K. 1991. "Artificial evolution for computer graphics." *Computer Graphics* 25: 319–28.

Stiny, G. and Mitchell, W. J. 1980. "The grammar of paradise: on the generation of Mughul gardens." *Environment and Planning B* 7: 209–26.

Walton, K. 1970. "Categories of art." *Philosophical Review* 79: 334–67.

Part III

Mind and AI

Chapter 9

The Philosophy of AI and its Critique

James H. Fetzer

Historical Background

Prior to the advent of computing machines, theorizing about the nature of mentality and thought was predominantly the province of philosophers, among whom perhaps the most influential historically has been René Descartes (1596–1650), often called "the father of modern philosophy." Descartes advanced an *ontic* (or ontological) thesis about the kind of thing minds are as features of the world and an *epistemic* (or epistemological) thesis about how things of that kind could be known. According to Descartes, who advocated a form of dualism for which mind and body are mutually exclusive categories, "minds" are things that can think, where access to minds can be secured by means of a faculty known as "introspection," which is a kind of inward perception of a person's own mental states.

Descartes's approach exerted enormous influence well into the twentieth century, when the development of digital computers began to captivate the imagination of those who sought a more scientific and less subjective conception of the nature of thinking things. The most important innovations were introduced by Alan Turing (1912–54), a brilliant British mathematician, cryptographer, theoretician, and philosopher. Some of Turing's most important research concerned the limitations of proof within mathematics, where he proposed that the boundaries of the computable (of mathematical problems whose solutions were obtainable on the basis of finite applications of logical rules) were the same as those that can be solved using a specific kind of problem-solving machinery.

Things of this kind, which are known as *Turing machines*, consist of an arbitrarily long segmented tape and a device capable of four operations upon that tape, namely: making a mark, removing a mark, moving the tape one segment forward, and moving the tape one segment backward. (The state of the tape before a series of operations is applied can be referred to as "input," the state of the tape after it has been applied as "output," and the series of instructions as a "program.") From the perspective of these machines, it became obvious there are mathematical problems for which no finite or computable solutions exist. Similar results relating effective procedures to computable problems were concurrently obtained by the great American logician Alonzo Church.

The Turing Test

Church's work was based on purely mathematical assumptions, while Turing's work appealed to a very specific kind of machine, which provided an abstract model for the physical embodiment of the procedures that suitably define "(digital) computers" and laid the foundation for the theory of computing. Turing argued that such procedures impose limits upon human thought, thereby combining the concept of a program with that of a mind in the form of a machine which in principle could be capable of having many types of physical implementation. His work thus introduced what has come to be known as *the computational conception of the mind*, which inverts the Cartesian account of machines as mindless by turning minds themselves into special kinds of machines, where the boundaries of computability define the boundaries of thought.

Turing's claim to have fathered AI rests upon the introduction of what is known as *the Turing test*, where a thing or things of one kind are pitted against a thing or things of another kind. Adapting a party game where a man and a woman might compete to see whether the man could deceive a contestant into mistaking him for the woman (in a context that would not give the game away), he proposed pitting a human being against an inanimate machine (equipped with a suitable program and mode of communication). Thus, if an interlocutor could not differentiate between them on the basis of the answers they provided to questions that they were asked, then those systems should be regarded as equal (or equipotent) with respect to (what he took to be) *intelligence* (Turing 1950).

This represented a remarkable advance over Cartesian conceptions in three different respects. First, it improved upon the vague notion of a thinking thing by introducing the precise notion of a Turing machine as a device capable of mark manipulation under the control of a program. Second, it implied a solution to the mind/body problem, according to which hardware is to software as bodies are to minds, that was less metaphorical and more scientific than the notion of bodies with minds. Third, it appealed to a behavioral rather than introspective criterion for empirical evidence supporting inferences to the existence of thinking things, making the study of the mind appear far less subjective.

Physical Machines

Descartes's conception of human minds as thinking things depends upon actually having thoughts, which might not be the case when they are unconscious (say, asleep, drugged, or otherwise incapable of thought), since their existence as things that think would not then be subject to introspective verification, which supports hypothesis (h1):

(h1) (Conscious) human minds are thinking things (Descartes);

Analogously, Turing's conception of these machines as thinking things depends upon the exercise of the capacity to manipulate marks as a sufficient condition for the possession of intelligence which could be comparable to that of humans, suggesting hypothesis (h2):

(h2) Turing machines manipulating marks possess intelligence (Turing);

where the identification of *intelligence* with *mentality* offers support for the conclusion that suitably programmed and properly functioning Turing machines might qualify as manmade thinking things or, in the phrase of John McCarthy, as "artificial intelligence."

As idealized devices that are endowed with properties that physical systems may not possess, including segmented tapes (or "memories") of arbitrary length and perfection in performance, however, Turing machines are *abstract entities.* Because they do not exist in space/time, they are incapable of exerting any causal influence upon things in space/time, even though, by definition, they perform exactly as intended (Fetzer 1988). The distinction is analogous to that between numbers and numerals, where numbers are abstract entities that do not exist in space/

time, while numerals that stand for them are physical things that do exist in space/time. Roman numerals, Arabic numerals, and such have specific locations at specific times, specific shapes and sizes, come into and go out of existence, none which is true of numbers as timeless and unchanging abstract entities.

These "machines," nevertheless, might be subject to at least partial implementations as physical things in different ways employing different materials, such as by means of digital sequences of 0s and 1s, of switches that are "on" or "off," or of higher and lower voltage. Some might be constructed out of vacuum tubes, others made of transistors or silicon chips. They then become instances of physical things with the finite properties of things of their kinds. None of them performs exactly as intended merely as a matter of definition: all of them have the potential for malfunction and variable performance like aircraft, automobiles, television sets, and other physical devices. Their memories are determined by specific physical properties, such as the size of their registers; and, while they may be enhanced by the addition of more memory, none of them is infinite.

Symbol Systems

While (some conceptions of) God might be advanced as exemplifying a timeless and unchanging thinking thing, the existence of entities of that kind falls beyond the scope of empirical and scientific inquiries. Indeed, within computer science, the most widely accepted and broadly influential adaptation of Turing's approach has been by means of *the physical symbol system conception* Alan Newell and Herbert Simon have advanced, where symbol systems are physical machines – possibly human – that process physical symbol structures through time (Newell & Simon 1976). These are special kinds of digital machines that qualify as serial processing (or von Neumann) machines. Thus, they implement Turing's conception by means of a physical machine hypothesis (h3),

(h3) Physical computers manipulating symbols are intelligent (Newell and Simon);

where, as for Turing, the phrase "intelligent thing" means the same as "thinking thing."

There is an ambiguity about the words "symbol systems" as systems that process symbols and as the systems of symbols which they process, where Newell and Simon focused more attention on the systems of symbols that machines process than they did upon the systems that process those symbols. But there can be no doubt that they took for granted that the systems that processed those symbols were physical. It therefore becomes important, from this point hence, to distinguish between "Turing machines" as abstract entities and "digital computers" as physical implementations of such machines, where digital computers, but not Turing machines, possess finite memories and potential to malfunction. Newell and Simon focused upon computers as physical machines, where they sought to clarify the status of the "marks" that computers subject to manipulation.

They interpreted them as sets of physical patterns they called "symbols," which can occur in components of other patterns they called "expressions" (or "symbol structures"). Relative to sets of alpha-numerical (alphabetical and numerical) characters (ASCII or EBCDIC, for example), expressions are sequences of symbols understood as sequences of characters. Their "symbol systems" as physical machines that manipulate symbols thus qualify as necessary and sufficient for intelligence, as formulated by hypothesis (h4):

(h4) (Being a) symbol system is both necessary and sufficient for intelligence (Newell and Simon);

which, even apart from the difference between Turing machines as abstractions and symbol systems as physical things, turns out to be a much stronger claim than (h2) or even (h3). Those hypotheses do not imply that every thinking thing has to be a digital computer or a Turing machine. (h2) and (h3) are both consistent with the existence of thinking things that are not digital computers or Turing machines. But (h4) does not allow for the existence of thinking things that are not digital machines.

The Chinese Room

The progression of hypotheses from (h1) to (h2) to (h3) and perhaps (h4) appears to provide significant improvement on Descartes's conception, especially when combined with the Turing test, since they not only clarify the nature of mind and elucidate the relation of mind to body, but even explain how the existence of other minds might be known, a powerful combination of ontic and epistemic theses that seems to support the prospects for artificial intelligence. As soon as computing machines were designed with performance capabilities comparable to those of human beings, it would be appropriate to ascribe to those inanimate entities the mental properties of thinking things. Or so it seemed, when the philosopher John Searle advanced a critique of the prospects for AI that has come to be known as "the Chinese Room" and cast it all in doubt (Searle 1980).

Searle proposed a thought experiment involving two persons, call them "C" and "D," one (C) fluent in Chinese, the other (D) not. Suppose C were locked in an enclosed room into which sequences of marks were sent on pieces of paper, to which C might respond by sending out other sequences of marks on other pieces of paper. If the marks sent in were questions in Chinese and the marks sent out were answers in Chinese, then it would certainly look as though the occupant of the room knew Chinese, as, indeed, by hypothesis, he does. But suppose instead D were locked in the same room with a table that allowed him to look up sequences of marks to send out in response to sequences of marks sent in. If he were very proficient at this activity, his performance might be the equal of that of C, who knows Chinese, even though D, by hypothesis, knows no Chinese.

Searle's argument was a devastating counterexample to the Turing test, which takes for granted that similarities in performance indicate similarities in intelligence. In the Chinese Room scenario, the same "inputs" yield the same "outputs," yet the processes or procedures that produce them are not the same. This suggests that a distinction has to be drawn between "simulations," where systems *simulate* one another when they yield the same outputs from the same inputs, and "replications," where systems *replicate* one another when they yield the same outputs from the same inputs by means of the same processes or procedures. In this language, Searle shows that, even if the Turing test is sufficient for comparisons of input/output behavior (simulations), it is not sufficient for comparisons of the processes or procedures that yield those outputs (replications).

Weak AI

The force of Searle's critique becomes apparent in asking which scenario, C or D, is more like the performance of a computer executing a program, which might be implemented as an automated look-up table: in response to inputs in the form of sequences of marks, a computer processes them into outputs in the form of other sequences of marks on the basis of its program. So it appears appropriate to extend the comparison to yet a third scenario, call it "E," where a suitably programmed computer takes the same inputs and yields the same outputs. For just as the performance of D might simulate the performance of C, even though D knows no Chinese, so the performance of E might simulate the performance of D, even though E possesses no mentality. Mere relations of simulation thus appear too weak to establish that systems are equal relative to their intelligence.

Searle also differentiated between what he called "strong AI" and "weak AI," where weak AI maintains that computers are useful tools in the study of the mind, especially in producing useful models (or simulations), but strong AI maintains that, when they are executing programs, computers properly qualify as minds (or replications). Weak AI thus represents an epistemic stance about the value of computer-based models or simulations, while strong AI represents an ontic stance about the kinds of things that actually are instances of minds. Presumably, strong AI implies weak AI, since actual instances of minds would be suitable subjects in the study of mind. Practically no one objects to weak AI, however, while strong AI remains controversial on many grounds.

That does not mean it lacks for passionate advocates. One of the most interesting introductions to artificial intelligence has been co-authored by Eugene Charniak and Drew McDermott (1985). Already in their first chapter, the authors define "artificial intelligence" as the study of mental faculties through the use of computational models. The tenability of this position, no doubt, depends upon the implied premise that mental faculties operate on the basis of computational processes, which, indeed, they render explicit by similarly postulating that what brains do "may be thought of at some level as a kind of computation" (Charniak & McDermott 1985: 6). The crucial distinction between "weak" and "strong" AI, however, depends upon whether brains actually qualify as computers, not whether they may be thought to be.

Strong AI

Charniak and McDermott also maintain "the ultimate goal of research in AI is to build a person or, more humbly, an animal." Their general conception is that the construction of these artificial things must capture key properties of their biological counterparts, at least with respect to kinds of input, kinds of processing, and kinds of output. Thus, the "inputs" they consider include vision (sights) and speech (sounds), which are processed by means of internal modules for learning, deduction, explanation, and planning, which entail search and sort mechanisms. These combine with speech and motor capabilities to yield "outputs" in the form of speech (sounds) and behavior (motions), sometimes called "robotics." The crucial issue thus becomes whether these "robots" are behaving like human beings as (mindless) simulations or instead embody (mindful) replications.

Their attention focuses upon what goes on in "the black box" between stimulus and response, where those with minds depend upon and utilize *internal representations* as states of such systems that describe or otherwise represent various aspects of the world. Indeed, some of these aspects could be internal to the system itself and thus represent its own internal states as internal representations of aspects of itself. But, while self-awareness and self-consciousness are often taken to be important kinds of intelligence or mentality, they do not appear to be essential to having intelligence or mentality in general as opposed to having intelligence or mentality of specific kinds. There may be various kinds of mentality or intelligence – mathematical, verbal, and artistic, for example – but presumably they share certain core or common properties.

There would seem to be scant room for doubt that, if artificial machines are going to qualify as comparable to human beings relative to their mental abilities, they must have the same or similar capacities to use and manipulate internal representations, at least with respect to some specified range – presumably, alpha-numeric – of tasks. They must take the same or similar external inputs (or "stimuli"), process them by means of the same or similar "mental" mechanisms, and produce the same or similar external outputs (or "responses"). While Charniak and McDermott may aspire to build an artificial animal, the AI community at large, no doubt, would settle for building an artificial thinking thing, presuming that it is possible to create one without the other.

Folk Psychology

There is an implied presumption that different systems that are subject to comparison are operating under the same or similar causally relevant background conditions. No one would suppose that a computer with a blown motherboard should yield the same outputs from the same inputs as a comparable computer with no hardware breakdown, even when they are loaded with the same programs. Analogously, no one would assume that a human being with a broken arm, for example, should display the same behavior in response to the same stimuli (say, a ball coming straight toward him while seated in the bleachers at a game) as another person without a broken arm. But that does not mean that they are not processing similar stimuli by means of similar representations.

Human beings are complicated mechanisms, whether or not they properly qualify as

"machines" in the sense that matters to AI Indeed, the full range of causally relevant factors that make a difference to human behavior appears to include motives, beliefs, ethics, abilities, capabilities, and opportunities (Fetzer 1996). Different persons with the same or similar motives and beliefs, for example, but who differ in their morals, may be expected to display different behavior under conditions where ethics makes a difference, even though they may have similar abilities and are not incapacitated from the exercise of those abilities. As we all know, human beings consume endless hours endeavoring to explain and predict the behavior of others and themselves, employing a framework of causally relevant factors of this kind, which has come to be known as "folk psychology."

No doubt when appraised from the perspective of, say, the conditions of adequacy for scientific theories – such as clarity and precision of language, scope of application for explanation and prediction, degree of empirical support, and the economy, simplicity, or elegance with which these results are attained – folk psychology appears to enjoy a high degree of empirical support by virtue of its capacity to subsume a broad range of cases within the scope of its principles. Some of that apparent success, however, may be due to the somewhat vague and imprecise character of the language upon which it depends, where there would appear to be opportunity for revision and refinement to enhance or confine its scope of application. Yet some researchers argue for its elimination altogether.

Eliminative Materialism

Paul Churchland, for example, maintains that folk psychology is not only incomplete but also inaccurate as a "misrepresentation" of our internal states and mental activities. He goes so far as to suggest that progress in neuroscience should lead, not simply to the refinement of folk psychology, but to its wholesale elimination (Churchland 1984: 43). The model Churchland embraces thus follows the pattern of elimination of "phlogiston" from the language of chemistry and of "witches" from the language of psychology. He thus contends that the categories of *motives* and *beliefs*, among others, are destined for a similar fate as neuroscience develops. Churchland admits he cannot guarantee that this will occur, where the history of science in this instance might instead simply reflect some adjustment in folk-psychological principles or dispensing with some of its concepts.

The deeper problem that confronts eliminative materialism, however, appears to be the same problem confronting classic forms of reductionism, namely, that without access to information relating brain states to mind states, on the one hand, and mind states to behavioral effects, on the other, it would be impossible to derive predictive inferences from brain states to behavioral effects. If those behavioral effects are manifestations of dispositions toward behavior under specific conditions, moreover, then it seems unlikely that a "mature" neuroscience could accomplish its goals if it lacked the capacity to relate brain states to behavioral effects by way of dispositions, because there would then be no foundation for relating mind states to brain states and brain states to human behavior.

In the case of jealousy (hostility, insincerity, and so on) as causal factors that affect our behavior in the folk-psychological scheme of things, if we want to discover the brain states that underlie these mind states as dispositions to act jealous (to act hostile, and so forth) under specific conditions, which include our other internal states, then a rigorous science of human behavior might be developed by searching for and discovering some underlying brain states, where those dispositions toward behavior were appropriately (presumably, lawfully) related to those brain states. Sometimes brain states can have effects upon human behavior that are not mediated by mind states, as in the case of brain damage or mental retardation. For neurologically normal subjects, mind states are able to establish connections between brain states and their influence on behavior.

Processing Syntax

The predominant approach among philosophers eager to exploit the resources provided by the

computational conception, however, has been in the direction of refining what it takes to have a mind rather than the relationship between minds, bodies, and behavior. While acknowledging these connections are essential to the adequacy of any account, they have focused primarily upon the prospect that language and mentality might be adequately characterized on the basis of purely formal distinctions of the general kind required by Turing machines – the physical shapes, sizes, and relative locations of the marks they manipulate – when interpreted as the alpha-numeric characters that make up words, sentences, and other combinations of sentences as elements of a language.

Jerry Fodor, for example, has observed that computational conceptions of language and mentality entail the thesis that "mental processes have access only to formal (nonsemantic) properties of the mental representations over which they are defined" (Fodor 1980: 307). He elaborates upon the relationship between the form (syntax) and the content (semantics) of thoughts, maintaining (a) that thoughts are distinct in content only if they can be identified with distinct representations, but without offering an explanation of how it is (b) that any specific thoughts can be identified with any specific representations, a problem for which he elsewhere offers a solution known as "the language of thought." But any account maintaining that the same syntax always has the same semantics or that the same semantics always has the same syntax runs afoul of problems with ambiguity on the one hand, and with synonymy on the other.

Nevertheless, the strongest versions of computational conceptions tend to eschew concern for semantics and focus instead on the centrality of syntax. Stephen Stich has introduced *the syntactic theory of the mind* (*STM*) as having an agnostic position on content, neither insisting that syntactic state types (as repeatable patterns of syntax) have no content nor insisting that syntactic state tokens (specific instances of syntactic state types) have no content: "It is simply silent on the whole matter . . . [T]he STM is in effect claiming that psychological theories have no need to postulate content or other semantic properties" (Stich 1983: 186). STM is thereby committed to hypothesis (h5):

(h5) Physical computers processing syntax possess minds (STM);

which may initially appear much stronger than (h3). But Newell and Simon's notion of "symbol" is defined formally and their "symbol systems" are also computing machines. Both approaches run the risk of identifying "thinking things" with mindless machines.

Semantic Engines

Systems of marks with rules for their manipulation are examples of (what are known as) *formal systems*, the study of which falls within the domain of pure mathematics. When those formal systems are subject to interpretations, especially with respect to properties and objects within the physical world, their study falls within the domain of applied mathematics. A debate has raged within computer science over whether that discipline should model itself after pure or applied mathematics (Colburn et al. 1993). But whatever the merits of the sides to that dispute, there can be scant room for doubt that mere mark manipulation, even in the guise of syntax processing, is not enough for thinking things. Thoughts possess content as well as form, where it is no stretch of the imagination to suggest that, regarding thought, content dominates form.

The STM, which makes syntax processing sufficient for the possession of mentality, thus appears to be far too strong, but a weaker version might still be true. The ability to process syntax might be necessary for mentality instead, as indeed hypothesis (h3) implies, when Newell and Simon's "symbols" are properly understood as marks subject to manipulation. Thus, a more plausible version of (h5) should maintain instead (h6):

(h6) (Conscious) minds are physical computers processing syntax;

where syntax consists of marks and rules for their manipulation that satisfy constraints that make them meaningful. But since there are infinitely

many possible interpretations of any finite sequence of marks, some specific interpretation (or class of interpretations) requires specification as "the intended interpretation." Marks can only qualify as syntax relative to specific interpretations in relation to which those marks become meaningful.

From this point of view, a (properly functioning) computing machine can be qualified as *an automatic formal system* when it is executing a program, but becomes meaningful only when its syntax satisfies the constraints of an intended interpretation. Indeed, an automatic formal system where "the semantics follows the syntax" has been designated "a semantic engine" by Daniel Dennett. This supports the contention some have called the basic idea of cognitive science: *that intelligent beings are semantic engines*, that is, automatic formal systems under which they consistently make sense (Haugeland 1981: 31). (h6) thus requires qualification to incorporate the role of interpretation as (h7):

(h7) Semantic engines are necessary and sufficient for intelligence;

where, as in the case of Newell and Simon, "intelligent things" are also "thinking things" and "(conscious) minds," understood as physical computers processing syntax under an interpretation. The problem is to "pair up" the syntax and the semantics the right way.

The Language of Thought

Jerry Fodor (1975) has advanced an argument hypothesizing the existence of an innate language, which is species-specific and possessed by every neurologically normal human being. He calls it *mentalese* (or "the language of thought"). He contends the only way to learn a language is to learn the truth conditions for sentences that occur in that language: "learning (a language) **L** involves learning that *'Px' is true if and only if x is G* for all substitution instances. But notice that learning that could be learning P (learning what P means) only for an organism that already understood G" (Fodor 1975: 80). Given the unpalatable choice between *an endless hierarchy* of successively richer and richer metalanguages for specifying the meaning of lower-level languages and *a base language* that is unlearned, Fodor opts for the existence of an innate and inborn language of thought.

The process of relating a learned language to the language of thought turns human beings into semantic engines, which may be rendered by hypothesis (h8) as follows:

(h8) Human beings are semantic engines with a language of thought (Fodor).

Fodor commits a mistake in his argument, however, by overlooking the possibility that the kind of prior understanding which is presupposed by language learning might be *nonlinguistic*. Children learn to suck nipples, play with balls, and draw with crayons long before they know that what they are doing involves "nipples," "balls," or "crayons." Through a process of interaction with things of those kinds, they acquire habits of action and habits of mind concerning the actual and potential behavior of things of those kinds. Habits of action and habits of mind that obtain for various kinds of things are concepts. Once that nonlinguistic understanding has been acquired, the acquisition of linguistic dispositions to describe them appears to be relatively unproblematical (Fetzer 1990).

One of the remarkable features of Fodor's conception is that the innate and inborn language of thought possesses a semantic richness such that this base language has to be sufficiently complete to sustain correlations between any natural language (French, German, Swahili, and such) at any stage of historical development (past, present, and future). This means that mentalese not only has to supply a foundation for everyday words, such as "nipple," "ball," and "crayon" in English, for example, but also those for more advanced notions, such as "jet propulsion," "polio vaccine," and "color television," since otherwise the language of thought could not fulfill its intended role. Among the less plausible consequences of this conception turn out to be that, since every human has the same innate language, which has to be complete in each of its instantiations, unsuccessful translations between

different languages and the evolution of language across time are both impossible, in principle, which are difficult positions to defend.

Formal Systems

Fodor's approach represents an extension of the work of Noam Chomsky, who has long championed the conception of an innate syntax, both inborn and species-specific, to which Fodor has added a semantics. Much of Chomsky's work has been predicated upon a distinction between competence and performance, where differences between the grammatical behavior of different language users, which would otherwise be the same, must be accounted for by circumstantial differences, say, in physiological states or psychological context. In principle, every user of language possesses what might be described as *(unlimited) computational competence*, where infinitely many sentences can be constructed from a finite base by employing recursive procedures of the kind that were studied by Church and Turing in their classic work on effective procedures.

Fodor and Zenon Pylyshyn (1988) adopt conditions for the production of sentences by language users implying that the semantic content of syntactic wholes is a function of the semantic content of their syntactic parts as their *principle of the compositionality of meaning* and that molecular representations are functions of other molecular or atomic representations as a *principle of recursive generability*. These conditions are obvious counterparts of distinctions between structurally atomic and structurally molecular representations as a precondition for a language of thought that is modeled on formal systems, such as sentential calculus. The principles of those formal systems, automated or not, may or may not transfer from abstract to physical contexts, not least because physical systems, including digital machines, are limited in their capacities.

Turing machines with infinite tapes and infallible performance are clearly abstract idealizations compared to digital machines with finite memories that can malfunction. The physical properties of persons and computers are decidedly different than those of automated formal systems as another case of abstract idealization. By comparison, digital machines and human beings possess no more than *(limited) computational competence* (Fetzer 1992). The properties of formal systems – such as incompleteness for systems richer than first-order monadic logic, which Kurt Gödel established – that might be supposed to impose limits on mental processes and have attracted interest by such scholars as J. R. Lucas (1961) and Douglas Hofstadter (1979), appear to have slight relevance to understanding the nature of cognition. Formal systems are useful in modeling reasoning, but reasoning is a special case of thinking. And if we want to understand the nature of thinking, we have to study thinking things rather than the properties of formal systems. Thinking things and formal systems are not the same.

Mental Propensities

Roger Penrose has suggested that thinking may be a quantum phenomenon and thereby qualify as *non-algorithmic* (Penrose 1989: 437–9). The importance of this prospect is that algorithms are commonly understood as functions that map single values within some domain onto single values within some range. If mental processes are algorithmic (functions), then they must be deterministic, in the sense that the same mental-state cause (completely specified) invariably brings about the same mental-state effect or behavioral response. Since quantum phenomena are not deterministic, if mental phenomena are quantum processes, they are not functions – not even partial functions, for which, when single values within a domain happen to be specified, there exist single values in the corresponding range, but where some of the values in the domain and range of the relevant variables might not be specified.

Systems for which the presence or the absence of every property that makes a difference to an outcome is completely specified are said to be "closed," while those for which the presence or absence of some properties that make a difference to the outcome are unspecified are said to

be "open." The distinction between deterministic and (in this case) probabilistic causation is that, for closed systems, for *deterministic* causal processes, the same cause (or complete set of conditions) invariably (or with universal strength *u*) brings about the same effect, whereas for *probabilistic* causal processes, the same cause variably (with probabilistic strength *p*) brings about one or another effect within the same fixed class of possible outcomes. A polonium^{218} atom, for example, has a probability for decay during a 3.05 minute interval of 1/2.

The determination that a system, such as an atom of polonium^{218}, is or is not a closed system, of course, poses difficult epistemic problems, which are compounded in the case of human beings, precisely because they are vastly more complex causal systems. Moreover, probabilistic systems have to be distinguished from (what are called) *chaotic systems*, which are deterministic systems with "acute sensitivity to initial conditions," where the slightest change to those conditions can bring about previously unexpected effects. A tiny difference in hundreds of thousands of lines of code controlling a space probe, for example, consisting of the occurrence of only one wrong character, a single misplaced comma, caused Mariner 1, the first United States interplanetary spacecraft, to veer off course and then have to be destroyed.

The Frame Problem

Indeed, there appear to be at least three contexts in which probabilistic causation may matter to human behavior, namely: in processing sensory data into patterns of neural activation; in transitions between one pattern of activation and another; and in producing sounds and other movement as a behavioral response. Processes of all three kinds might be governed by probabilistic or by chaotic deterministic processes and therefore be more difficult to explain or predict, even when the kind of system under consideration happens to be known. The most important differences between species appear to concern the range and variety of sensory data they are capable of processing, the speed and reliabilty with which they can effect transitions between patterns of activation, and the plasticity and strength of their behavior responses.

Concerns about variation in such types of causation also arise within the context of the study of mental models or representations of the world, specifically, what has be known as *the frame problem*, which Charniak and McDermott describe as the need to infer explicitly that one or more states will not change across time, which forms a "frame" within which other states may change (1985: 418) While the frame problem has proven amenable to many different characterizations – a variety of which may be found, for example, in Ford and Hayes 1991 – one important aspect of the problem is the extent to which a knowledge base permits the prediction and the explanation of systems when those systems are not known to be open or closed.

Indeed, from this point of view, the frame problem even appears to instantiate the classic *problem of induction* encountered in attempting to predict the future based upon information about the past, which was identified by David Hume (1711–76), a celebrated Scottish philosopher. Thus, Hume observed that there are no deductive guarantees that the future will resemble the past, since it remains logically possible that, no matter how uniformly the occurrence of events of one kind have been associated with events of another, they may not continue to be. If the laws of nature persist through time, however, then, in the case of systems that are closed, it should be possible to predict – invariably or probabilistically – precisely how those systems will behave over intervals of time, so long as the complete initial conditions and laws of systems of that kind are known.

Minds and Brains

Because connectionism appeals to patterns of activation of neural nodes rather than to individual nodes as features of brains that function as representations and affect behavior, it appears to improve upon computationally-based conceptions in several important respects, including perceptual completions of familiar patterns by filling

in missing portions, the recognition of novel patterns even in relation to previously unfamiliar instances, the phenomenon known as "graceful degradation," and related manifestations of mentality (Rumelhart et al. 1986: 18–31). Among the most important differences is that connectionist "brains" are capable of what is known as *parallel processing*, which means that, unlike (sequential) Turing machines, they are capable of (concurrently) processing more than one stream of data at the same time.

This difference, of course, extends to physical computers, which can be arranged to process data simultaneously, but each of them itself remains a sequential processor. The advantages of parallel processing are considerable, especially from the point of view of evolution, where detecting the smells and the sounds of predators before encountering the sight of those predators, for example, would afford adaptive advantages. Moreover, learning generally can be understood as a process of increasing or decreasing activation thresholds for specific patterns of nodes, where classical and operant conditioning may be accommodated as processes that establish association between patterns of activation and make their occurrence, under similar stimulus conditions, more (or less) probable, where the activation of some patterns tends to bring about speech and other behavior.

Those who still want to defend computational conceptions might hold that, even if their internal representations are distributed, human beings are semantic engines (h9):

(h9) Human beings are semantic engines with distributed representations;

but the rationale for doing so becomes less and less plausible and the mechanism – more and more "independent but coordinated" serial processors, for example – appears more and more *ad hoc*. For reasons that arose in relation to eliminative materialism, however, no matter how successful connectionism as a theory of the brain, it cannot account for the relationship between bodies and minds without a defensible conception of the mind that should explain why symbol systems and semantic engines are not thinking things.

Semiotic Systems

The conception of minds *as semiotic (or as sign-using) systems* advances an alternative to computational accounts that appears to fit the connectionist model of the brain like a hand in a glove. It provides a noncomputational framework for investigating the nature of mind, the relation of mind to body, and the existence of other minds. According to this approach, *minds* are things for which something can stand for something else in some respect or other (Fetzer 1990, 1996). The semiotic relation, which was elaborated by the American philosopher Charles S. Peirce (1839–1914), is triadic (or three-placed), involving a relation of *causation* between signs and their users, a (crucial) relation of *grounding* between signs and that for which they stand, and an *interpretant* relation between signs, what they stand for, and the users of signs.

There are three branches of the theory of semiotics, which include *syntax* as the study of the relations between signs and how they can be combined to create new signs, *semantics* as the study of the relations between signs and that for which they stand, and *pragmatics* as the study of the relations between signs, what they stand for, and sign users. Different kinds of minds can then be classified on the basis of the kinds of signs they are able to utilize – such as *icons*, which resemble that for which they stand (similar in shapes, sizes, and such); *indices*, which are causes or effects of that for which they stand (ashes, fires, and smoke); and *symbols*, which are merely habitually associated with that for which they stand (words, sentences, and things) – as iconic, indexical, and symbolic varieties of mentality, respectively.

Meanings are identified with the totality of possible and actual behavior that a sign user might display in the presence of a sign as a function of context, which is the combination of motives, beliefs, ethics, abilities, and capabilities that sign-users bring to their encounters with signs. And patterns of neural activation can function as internal signs, where (all and only) thinking things are semiotic systems, (h10):

(h10) Thinking things, including human beings, are semiotic systems.

This approach can explain what it is to be conscious relative to a class of signs, where a system is *conscious* with respect to signs of that kind when it has the ability to utilize signs of that kind and is not inhibited from the exercise of that ability. And it supports the conception of *cognition* as an effect that is brought about (possibly probabilistically) by interaction between signs and sign-users when they are in suitable causal proximity.

Critical Differences

Among the most important differences between semiotic systems and computational accounts becomes apparent at this point, because the semantic dimension of mentality has been encompassed by the definition of systems of this kind. Observe, for example, the difference between symbol systems and semiotic systems in figures 9.1 and 9.2, where semiotic systems reflect a grounding relationship that symbol systems lack.

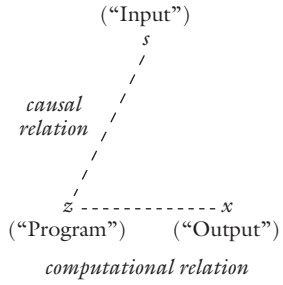

Figure 9.1: Symbol systems

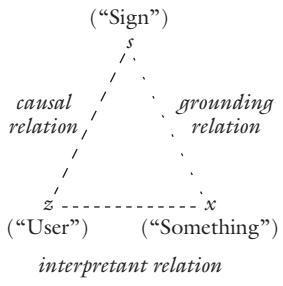

Figure 9.2: Semiotic systems

This difference applies even when these systems are processing marks by means of the same procedures. A computer processing a tax return can yield the same outputs from the same inputs, yet they mean nothing to that system as income, deductions, or taxes due. A distinction must be drawn between those marks that are meaningful for use by a system and marks that are meaningful for the users of that system. They can function as signs for those users without having to function as signs for those systems.

"Symbols" in this sense of semiotic systems must therefore be clearly distinguished from "symbols" in the sense of symbol systems, which can be meaningless marks, lest one mistake symbol systems in Newell and Simon's sense for (symbol-using) semiotic systems, as has John McCarthy (McCarthy 1996: ch. 12). This reflects (what might be called) *the static difference* between computer systems and thinking things. Another is that digital machines are under the control of programs as causal implementations of algorithms, where "algorithms" in turn are effective decision procedures. Effective decision procedures are completely reliable in producing solutions to problems within appropriate classes of cases that are invariably correct and they do in a finite number of steps. If these machines are under the control of algorithms but minds are not, then there is *a dynamic difference* that may be more subtle but is not less important as well.

Indeed, there are many kinds of thinking – from dreams and daydreams to memory and perception as well as ordinary thought – that do not satisfy the constraints imposed by effective decision procedures. They are not reliable problem-solving processes and need not yield definitive solutions to problems in a finite number of steps. The causal links that affect transitions between thoughts appear to be more dependent upon our life histories and associated emotions (our pragmatic contexts) than they do on syntax and semantics *per se*. Even the same sign, such as a red light at an intersection, can be taken as an icon (because it resembles other red lights), as an index (as a traffic control device that is malfunctioning), or as a symbol (where drivers should apply the breaks and come to a complete halt) as a function of a sign user's context at the time. Anyone else in the same context would

(probabilistically) have interpreted that sign the same way.

The Hermeneutic Critique

Whether or not the semiotic conception ultimately prevails, current research makes it increasingly apparent that an adequate account of mentality will have to satisfy many of the concerns raised by the hermeneutic critique advanced by Hubert Dreyfus (1979). Dreyfus not only objected to the atomicist conception of representation that became the foundation for the compositionality of meaning and recursive generability theses that Fodor and Pylyshyn embraced but also emphasized the importance of the role of bodies as vehicles of meaning, especially through interactions with the world, very much in the spirit of Peirce, which whom he shares much in common. Thus, the very idea of creating artificial thinking things whose minds are not inextricably intertwined with their bodies and capable of interacting with the world across time becomes increasingly implausible.

It has become clear that differences between Turing machines, digital computers, and human beings go considerably beyond those addressed above, where the semiotic conception of consciousness and cognition, for example, offers the capacity to make a mistake as a general criterion of mentality, where making a mistake involves taking something to stand for something else, but doing so wrongly, which is the right result. From this point of view, there appear to be three most important differences (see table 9.1). Even apart from a specific theory of representation intended to account for the meaning of the marks machines can manipulate, it appears evident from the table that these are three distinctly different kinds of things, where thinking things are unlike digital machines.

Ultimately, of course, the adequacy of a theory of mind hinges upon the adequacy of the theory of meaning it provides that relates brains, minds, and behavior. The crucial consideration appears to be that, whether bodies and minds are deterministic, chaotic, or probabilistic systems, it must provide a completely causal account of how the signs that minds employ make a difference to the behavior of those systems that is sufficient to sustain an inference to the existence of mentality as the best explanation for the data. One way in which that may occur emerges from the different ways in which sensations affect behavior, where the dog barked at the bush when the wind blew, because he mistook it for a stranger; where Mary rushed to the door at the sound of the knock, because she thought her friend had come; or where Bob slowed down when the light turned red, because he knew that he should apply the breaks and bring the car to a complete halt.

Conventions and Communication

Because different users can use different signs with the same meaning and the same signs with different meaning, it is even possible for a sign user to use signs in ways that, in their totality, are not the same as those of any other user. This implies that social conceptions of language, according to which private languages are impossible, are not well-founded from the perspective of semiotic systems. A person who found himself abandoned on a deserted island, for example, might while away the time by constructing an elaborate system of classification for its flora and fauna. Even though that system of signs might therefore have unique significance for that

Table 9.1: Three distinctly different kinds of things

	(Abstract) Turing machines	*(Physical) Digital computers*	*(Actual) Human beings*
Infinite capacities:	Yes	No	No
Subject to malfunction:	No	Yes	Yes
Capable of mistakes:	No	No	Yes

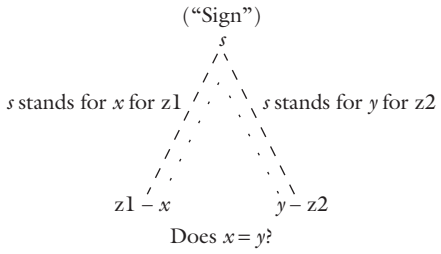

Figure 9.3: Communication situations

individual user, that system of signs, presumably, would still be learnable in the sense that there is no reason why it could not be taught to others. It would simply be the case it never had.

In communication situations, whether spoken, written, or otherwise, different sign users tend to succeed when they use signs the same way or to the extent to which they mean the same things by them. The question that arises is whether the same sign *s* stands for the same thing *x* for different sign users z1 and z2 under specific conditions (see figure 9.3). When z1 and z2 speak different languages, such as English and German, the success of a translation can be difficult to ascertain. But it can also be difficult when very similar sounds are associated with meanings that may not mean the same thing for every user.

There are circumstances under which we may prefer for our signs to be confidential. Turing himself, for example, spent time successfully cracking the Enigma cipher during the Second World War, enabling the British to understand the German's coded messages. Other circumstances, however, encourage the use of the same signs in the same ways, such as in the case of a community of members with common objectives and goals. Systems of public schools, for example, are commonly financed with the purpose, among others, of instilling the members of the community with a common understanding of the language they use, which promotes communication and cooperation between them. Some nations, such as the United States, have benefited immeasurably from their standing as "melting pots" where people from many countries come together and are united by reliance upon English, in the absence of which this country would no doubt tend toward Balkanization.

Other Minds

The adoption of an approach of this general kind clarifies and illuminates distinctively mental aspects of various sorts of causal processes. When causal relations occur (when causes such as inputs bring about effects such as outputs) and those inputs and outputs do not serve as signs for a system, they may then be classified as stimuli. When effects are brought about by virtue of their grounding (because they stand for those things in those respects) for the systems that use them, they may properly be classified as signs. And when semiotic relations occur (when signs being used by one user are interpreted by another) between systems that use them, they may be further classified as signals. Sometimes the signals we send are intentional (successful, and so on), sometimes not. Every sign must be a stimulus and every signal must also be a sign, but not vice versa.

Every human being, (other) animal, and inanimate machine capable of using signs thereby qualifies as a thinking thing on the semiotic conception. This realization thus explains why dreams and daydreams, memory and perception, and ordinary thought are mental activities, while tooth decay, prostate cancer, and free-fall, by comparison, are not. Whether or not the semiotic conception emerges as the most adequate among the alternative conceptions, it has become apparent that an adequate account ought to be one that is at least very much like it, especially in accommodating crucial differences between Turing machines, digital computers, and human beings. It has become equally apparent, I surmise, that minds are not machines. If thinking were governed by mental algorithms, as such accounts imply, then minds simply follow instructions mechanically, like robots, and have no need for insight, ingenuity, or invention. Perhaps we deny that we are nothing but robots because our mental activities involve so much more. Indeed, some of the most distinctive aspects of thought tend to separate minds from machines.

Simulations are clearly too weak and *emulations*, which yield the same inputs from the same outputs by means of the same processes and are made of the same matter, are clearly too strong. But the shoals are treacherous. David Chalmers,

for example, has argued that, for some systems, simulations are replications, on the presumption that the same psychophysical laws will be operative. Thus, if the transition from an initial state S_1 at time t_1 yields a final state S_n at t_n, where the intermediate steps involved in the transition between them, say, S_2 at t_2 through S_{n-1} at t_{n-1}, are the same, then properties that are lawfully related to them, such as consciousness, must come along with them, even when they are made of different stuff (Chalmers 1996). But that will be true only if the difference in matter does not affect the operation of those laws themselves. In cases where it does, *replications may require emulations.*

Intelligent Machines

An approach of this kind can explain why symbol systems and semantic engines are not thinking things. Their properties account for the form of thoughts but not their content, or else cannot account for the transitions between thoughts themselves. Turing machines, with which we began, are not even physical things and cannot sustain the existence of finite minds that can malfunction and can make mistakes. The connectionist conception of brains as (wet) neural networks supplies a crucial foundation for rethinking the nature of the mind, but requires supplementation by an account of the nature of the mind that is *noncomputational.* An appropriate conception of mental causation – including the processes of perception, of thought transition, and of response behavior – should permit those kinds of processes to be computational but not require it. Computing is merely one special kind of thinking.

Not the least of the benefits that are thereby derived is an account of mentality that can be reconciled with biology and evolution. Primitive organisms must have had extremely elementary semiotic abilities, such as sensitivity to light by means of single cells with flagella to bring about motion. If moving toward the light promotes survival and reproduction, then that behavior would have adaptive benefits for such simple systems. Under the combined influence of genetic mutation, natural selection, genetic drift, sexual reproduction, sexual selection, group selection, artificial selection and genetic engineering, of course, biological evolution, including of our own species, continues to this day, bringing about more complex forms of semiotic systems with abilities to use more signs of similar kinds and other signs of various different kinds.

As manmade connectionist systems of (dry) neural networks are developed, it should not be too surprising if they reach a point where they can be appropriately classified as *artificial thinking things.* Whether that point will ever come depends upon advances in science and technology over which philosophers have no control. While the conception of symbol systems and even semantic engines appear to fall short of capturing the character of thinking things, this does not mean that they fail to capture the character of intelligent machines. To the extent to which machines properly qualify as "intelligent" when they have the ability to process complex tasks in a reliable fashion, the advent of intelligent machines came long ago. The seductive conceptual temptation has been to confuse intelligent machines with thinking things.

See also Chapter 1, COMPUTATION; Chapter 2, COMPLEXITY; and especially Chapter 10, COMPUTATIONALISM, CONNECTIONISM, AND THE PHILOSOPHY OF MIND.

References

Chalmers, D. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* New York, NY: Oxford University Press. [One of the most sophisticated discussions of the nature of mind by a leading representative of the computational conception. Intermediate.]

Charniak, E. and McDermott, D. 1985. *Introduction to Artificial Intelligence.* Reading, MA: Addison-Wesley. [An encompassing and sophisticated introduction to this discipline. Introductory to advanced.]

Churchland, P. 1984. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind.* Cambridge, MA: MIT Press. [A lucid discussion of the philosophy of mind and AI with

emphasis on eliminativism and the relationship of minds and bodies. Introductory.]

Colburn, T. et al., eds. 1993. *Program Verification: Fundamental Issues in Computer Science*. Dordrecht, the Netherlands: Kluwer Academic. [A collection of the most important papers. Level varies.]

Dreyfus, H. 1979. *What Computers Still Can't Do: A Critique of Artificial Reason*, rev. ed. New York: Harper & Row. [A critique of the foundations of AI from multiple philosophical perspectives. Intermediate.]

Fetzer, J. H. 1988. "Program verification: the very idea." *Communications of the ACM* 31: 1048–63. [A study of the applicability of formal methods within computer science that ignited a controversy in the field. Advanced.]

——. 1990. *Artificial Intelligence: Its Scope and Limits*. Dordrecht, the Netherlands: Kluwer Academic. [A sustained study of the theoretical foundations of artificial intelligence. Intermediate to advanced.]

——. 1992. "Connectionism and cognition: why Fodor and Pylyshyn are wrong." In A. Clark and R. Lutz, eds., *Connectionism in Context*. Berlin: Springer-Verlag, pp. 37–56.[A critique of Fodor and Pylyshyn's attempt to reject connectionism on formal-system principles. Intermediate.]

——. 1996. *Philosophy and Cognitive Science*, 2nd ed. St. Paul, MN: Paragon House. [An introduction to cognitive science. Introductory.]

Fodor, J. 1975. *The Language of Thought*. Cambridge, MA: MIT Press. [An influential position that both fascinates and infuriates. Intermediate.]

——. 1980. "Methodological solipsism as a research strategy in cognitive psychology." In J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: The MIT Press, pp. 307–38. [Interesting reflections on methodology. Intermediate.]

—— and Pylyshyn, Z. 1988. "Connectionism and cognitive architecture: a critical analysis." *Cognition* 28: 3–71. [A criticism of connectionism rooted in a formal systems conception of language and mentality. Advanced.]

Ford, K. M. and Hayes, P., eds. 1991. *Reasoning Agents in a Dynamic World: The Frame Problem*. Greenwich, CT: JAI Press. [A broad range of conceptions of the frame problem are presented and explored. Intermediate to advanced.]

Haugeland, J. 1981. "Semantic Engines: an introduction to mind design." In J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: MIT Press, pp. 1–34. [A brilliant introduction to a popular and influential volume. Introductory.]

Hofstadter, D. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books. [A Pulitzer Prize winning study of art, music, and mathematics which explores the potential for AI. Introductory to advanced.]

Lucas, J. R. 1961. "Minds, machines, and Gödel." *Philosophy* 36: 112–27. [Deploys Gödel against the conception of minds as machines. Advanced.]

McCarthy, J. 1996. *Defending AI Research*. Stanford: CSLI Lecture Notes. [A collection of essays, principally book reviews, in which one of the original contributors to AI explains why he disagrees with other views. Intermediate.]

Newell, A. and Simon, H. 1976. "Computer science as empirical inquiry: symbols and search." In J. Haugeland, ed., *Mind Design: Philosophy, Psychology, Artificial Intelligence*. Cambridge, MA: MIT Press, 1981, pp. 35–66. [A classic paper that has exerted great influence among computer scientists and deserves more study from philosophers. Intermediate.]

Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press. [An attack on the notion that all of human thought is algorithmic. Intermediate.]

Rumelhart, D. et al. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vols. 1 (*Foundations*) and 2 (*Psychological and Biological Models*). Cambridge, MA: MIT Press. [The studies that introduced neural networks, with great sensitivity to philosophical implications. Intermediate to advanced.]

Searle, J. 1980. "Minds, brains, and programs." *Behavior and Brain Sciences* 3: 417–57. [This classic paper provides a systematic response to Turing (1950), including replies to Turing's responses to criticism. Introductory.]

Stich, S. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press. [A brilliant analysis of alternative accounts of the mind that challenges ordinary folk psychology. Intermediate.]

Turing, A. M. 1950. "Computing machinery and intelligence." *Mind* LIX: 433–60. [The classic paper in which Turing introduces his comparative test, including analysis of eight lines of objection to his test. Introductory to intermediate.]

Chapter 10

Computationalism, Connectionism, and the Philosophy of Mind

Brian P. McLaughlin

Introduction

The central questions of the philosophy of mind are the nature of mental phenomena, and how mental phenomena fit into the causal structure of reality. The computational theory of mind aims to answer these questions. The central tenet of the theory is that a mind is a computer. According to the theory, mental states and events enter into causal relations via operations of the computer. The main aim of the theory is to say what kind of computer – what kind of computational mechanism – a mind is. The answer is still unknown. Pursuing it is the main research program of the theory.

In the most general sense, a computer is, roughly, a system of structures functionally organized in such a way as to be able to compute. The structures, their functional organization, and the basic modes of operation of the system when it computes comprise the functional architecture of the computer. The two tasks of the computational theory of mind are: (1) to identify the functional architecture of the computing system that grounds our mental abilities and (2) to explain how those abilities are exercised via operations of the system. The tasks are related. The explanation of how operations of the system constitute exercises of our mental abilities will justify the claim that our possession of those abilities consists in our being at least partly constituted by the system.

Computationalists hold that the functional architecture of the computing system that grounds our mental abilities resides in our brains. There is, however, no consensus as to what even the general character of that architecture is. The symbols-system paradigm and the connectionist paradigm are the two dominant research paradigms within the computational theory of mind. They differ primarily in what kind of computer the mind is assumed to be, and thus in the kinds of functional architectures explored. The symbol-system paradigm presupposes that the mind is a kind of automatic formal system, while the connectionist paradigm presupposes that it is a system of connectionist networks. These paradigms will be discussed in due course. First, however, further general discussion of the computational theory of mind is in order.

1 The Computational Theory of Mind

Cognitive functions

Having a mind consists in having mental abilities such as, for example, the ability to think. According to the computational theory of mind, the exercise of mental abilities consists in the computation of certain functions – cognitive functions. The theory is thus concerned with what functions are cognitive functions and how they are computed.

Functions are relations, but not every relation is a function. A function is a one–one or many–one relation. Functions have arguments and values. The set of arguments of the function is its domain and the set of values is its range. For each argument in the domain, there is a unique value in the range. Functions are thus extensionally characterized as sets of ordered pairs, the first element of each ordered pair being an argument of the function, the second, the unique value of the function for that argument. Addition, multiplication, and division, for example, are mathematical functions that take order pairs of numbers as arguments and have numbers as values. A computable function is one that can be computed. Not all functions are computable. A function is computable if and only if there is an algorithm for finding a value of the function for any argument of the function. An algorithm is a kind of procedure, a way of getting something done; but not every procedure is an algorithm. An algorithm is an effective procedure, a sure-fire way of getting something done in a finite number of steps. It is sure-fire in that if each step is followed exactly, success is guaranteed. Each step of an algorithm is, moreover, easy in the sense that it requires no ingenuity or intelligence to carry it out, and thus "a mere mechanism" could execute it.

To compute is to compute a function, and to compute a function is to execute an algorithm. More than one algorithm can be used to compute a function. For example, the multiplication function – understood extensionally as a set of ordered pairs – can be computed using either the partial-product method taught in elementary school or by a series of additions. In fact, for any computable function, there will be infinitely many algorithms for computing it. The reason is trivial: we can always add intermediate steps to an algorithm for computing the function that contribute nothing to its computation.

How a function is computed will depend on the functional architecture of the machine (the computing system) that computes it, including the representational code used in the machine. Thus, for example, the algorithms for doing addition in base 10 differ from those for doing addition in binary notion, which uses only 1's and 0's. Likewise, the algorithms for doing addition, multiplication, and division using Roman numerals are different from those for computing these functions in the Arabic numerals we were taught in grade school.

These methods of calculating are algorithms for manipulating numerals (symbols that represent numbers), as are the methods for calculation using other kinds of numeral systems. However, the symbols manipulated by an algorithm can be symbols for anything that can be represented: persons, places, things, etc. And since a symbol can purport to represent something yet fail, an algorithm can manipulate symbols that purport to represent something yet fail.

Symbols have formal as well as semantic properties. The formal properties of symbols are intrinsic nonsemantic properties, such as shape or syntactic structure. Symbolic algorithms operate on symbols in virtue of certain of their formal properties, rather than their semantic ones. Thus, for example, the mathematical operations on Arabic numerals are defined over the shapes of the numerals, and likewise for the mathematical operations on Roman numerals. The beauty of symbolic algorithms is that they can involve the manipulation of symbols in ways that makes sense given what the symbols represent. Thus, while the algorithms for adding Arabic numerals are different from those for adding Roman numerals, the symbol transitions that occur during the execution of the algorithms in question can be systematically interpreted as adding the numbers represented.

Not all algorithms are symbol manipulation algorithms. An algorithm for getting all of the sides of a Rubik's Cube to match in color, for

example, involves the manipulation of the squares of the Cube, not the manipulation of symbols representing them (Cummins & Schwarz 1991). The computational theory of mind is concerned with the algorithms by which we compute cognitive functions. As will become apparent in due course, while the symbol-system paradigm appeals to symbol manipulation algorithms to explain how cognitive functions are computed, the connectionist computational paradigm often appeals to nonsymbolic algorithms to do so.

When we execute the partial-product algorithm using pencil and paper, for instance, each step of the process is guided by our intentions and decisions; likewise, when we execute by hand an algorithm for getting all of the sides of a Rubik's Cube to match in color. Computationalists maintain that the execution of basic cognitive algorithms (ones not executed by means of the execution of other cognitive algorithms) can be described and explained without invoking intelligent guidance for any step, and so without having to invoke "homunculi" – little intentional agents. This ultimate elimination of homunculi is essential since computationalism aims to reductively explain intentional abilities. (See the discussion of homunculi in Sterelny 1990.)

To see how an algorithm can be executed completely mindlessly, consider electronic circuits that function as truth-functional gates. For example, an AND-gate has two input wires and one output wire. When it receives current along both of its input wires it closes, thereby sending current along its output wire. When it receives current from only one input wire or does not receive current from either input wire, it remains open and no current is sent along its output wire. This circuit is called an "AND-gate" because using *T* to represent current being sent current along a wire and *F* to represent the absence of current being sent along a wire, we can construct a machine table representing how this circuit works that will exactly resemble the truth-table for conjunction, the truth function expressed by central uses of "and" in English. An AND-gate can thus be interpreted as computing the truth-function conjunction. Indeed for any truth table, no matter how complex, there is a formally equivalent electronic circuit design, and the converse. The electronic circuitry of such a device acts completely mindlessly.

The computational, algorithmic, and implementational levels

David Marr (1982) perspicuously distinguished three levels of analysis for addressing a computational problem. At the computational level of analysis, one specifies *what* cognitive function is being computed. At the algorithmic level, one describes *how* the function is being computed, the algorithm used to compute it. And at the implementational level, one describes how the steps of that algorithm are implemented, that is, the underlying mechanism by whose operations they are taken.

The three levels are relative to a computational problem. What is at the implementational level relative to one computational problem can be at either the computational or algorithmic level relative to another. The reason is that the implementation of a step of an algorithm might itself be characterized as the computation of a function, one executed by a different algorithm, whose steps are themselves implemented somehow; and each step of the latter algorithm might itself be characterized as the computation of a function, and so on, and so forth. Different algorithms are executed at different scales in the brain. It is an unanswered empirical question whether all cognitive functions are executed by algorithms involving elements at the same scale (Churchland & Sjenowski 1993).

Marr (1982) suggested that computational level analysis could be carried out largely independently of algorithmic level analysis, and the latter largely independently of implementational analysis. Now it is certainly true that the computational level underdetermines the algorithmic level: infinitely many algorithms can compute the same function. Moreover, the algorithmic level in turn underdetermines the implementational level: a given algorithm can be implemented in infinitely many possible ways. Nonetheless, the computationalist's concern is not merely to discover what cognitive functions are computed, but also to discover the algorithms used to compute them. Since what algorithms a machine can

execute will depend on its functional architecture, attention to likely modes of implementation can help in discovering those algorithms.

What algorithms a machine can execute supervenes on the functional architecture of the machine. (*A*-properties supervene on *B*-properties just in case the two things cannot differ in respect of *A*-properties without differing in respect of *B*-properties.) Thus, there cannot be a difference in what algorithms two machines can compute without a difference in their functional architectures. It follows that two machines with exactly the same functional architecture can execute exactly the same algorithms. Similarly, what functions a machine can compute will supervene on what algorithms it can execute, and so two machines that can execute exactly the same algorithms can compute exactly the same functions. Since supervenience is transitive, what functions a machine can compute will supervene on the functional architecture of a machine. As indicated above, however, functional architecture does not supervene on computational power. Machines with very different architectures can have exactly the same computational power.

Turing machines and Turing computability

In his early machine-state psychofunctionalism, Hilary Putnam (1975) claimed that we are finite state automata, by which he meant that we are Turing machines with finite tapes. The idea that our minds have the functional architecture of a Turing machine with a finite tape is not to be taken seriously. No computationalist thinks the mind has a Turing machine architecture. One of Putnam's essential points, however, was that we have the computational capacities of a Turing machine with a finite tape, in that our cognitive functions are computable by a Turing machine. According to the computational theory of mind, all cognitive functions can be computed by algorithms – effective procedures. According to the Church–Turing thesis, every function that can be computed by an algorithm can be computed by a Turing machine. If the Church–Turing thesis is correct, then the computational theory of mind is committed to the thesis that every cognitive function is Turing-computable (see Chapter 1, Computation).

Since not all functions are computable by a Turing machine, a way to try to falsify the computational theory of mind is by showing that there are cognitive functions that are not so computable. Whether there are such cognitive functions is a subject of dispute, and so this remains one line of attack on computationalism (Lucas 1961; Penrose 1989; van Gelder 1995; McCall 1999; Copeland 2000; see Lewis 1979 for a response to Lucas; and see Chalmers 1996 for responses to Penrose).

The Holy Grail of artificial intelligence

Since the functional architecture of our minds resides in our brains, Marvin Minsky has called our minds "meat machines." The Holy Grail of the field of artificial intelligence (see Chapter 9, The Philosophy of AI and its Critique) is to build something with mental abilities out of something other than living tissue. In pursuit of this Holy Grail, AI presupposes the computational theory of mind. If the computational theory of mind is correct, then it is at least logically possible for a mind essentially like ours to be made of quite different stuff from ours. The reason is that what is essential are the structures and their functional organization, not the material of which the elements of the structures are made. If the computational theory is right, then it is at least logically possible that the relevant structures with the relevant functional organization can be realized in something that is, for example, silicon based, rather than carbon based.

AI has been impressively successful in designing machines that can perform difficult tasks without our supervision. Moreover, it has been at the cutting edge of research into how cognitive functions might be computed. Finding the Holy Grail, however, remains an unrealized dream. We have computers that play master's-level chess and teams of robots that play soccer, but there are no artifacts that are even remotely plausible candidates for being the subjects of mental abilities. And despite the optimism of some AI researchers, there do not seem to be any in the works. It is

important to note, then, that the computational theory of mind is not committed to the success of this AI project. It leaves it an open empirical question whether the dream will ever be realized. For all we know, it may be that given the laws of nature and the initial conditions of the universe it is physically impossible for a mind (one essentially like ours) to be realized in something composed entirely of (e.g.) silicon. Such a mind might be impossible in the way that a 90-foot tall human being is impossible. The computational theory leaves this empirical issue open.

Nor is the computational theory of mind committed to Turing's (1950) would-be test (see Chapter 9) for determining whether something is genuinely intelligent (as opposed to merely appearing intelligent). The theory leaves open that there is more to intelligence than matching the dispositions to verbal or written behavior of an intelligent human being. It is not committed to the view that there is some pattern of dispositions to outward or peripheral behavior that suffices for the possession of genuine intelligence (Block 1981; Jackson 1993; McLaughlin & O'Leary-Hawthorne 1994).

2 The Symbol-system Paradigm

Mind as an interpreted automatic formal system

According to proponents of the symbol-system paradigm, the mind is an interpreted automatic formal system (Haugeland 1985). More precisely, having mental abilities consists in being constituted, at least partly, by an interpreted automatic formal system. A formal system is a system for manipulating discrete items in virtue of their formal properties; an automatic formal system is one that does so automatically. Many games, among them chess and checkers, are formal systems. Games that are formal systems share the following features: they are finitely playable, and are played by manipulating discrete items according to the rules of the game. In chess for instance, the discrete items are the chess pieces, which are manipulated according to the rules of chess. An automated formal system is one in which discrete items are automatically manipulated according to the rules – like a chess set that plays chess by itself. An interpreted automatic formal system is an automated formal system in which the discrete items that are manipulated include symbols or representations: discrete items with semantic and formal properties. The rules by which they are manipulated prescribe algorithmic operations on them. Thus, this paradigm is sometimes called "the rules and representations" paradigm.

Turing machines are interpreted automatic formal systems. There are many computationally equivalent machines that are automatic formal systems with different functional architectures from that of a Turing machine (see Chapter 1, COMPUTATION). One such machine is a von Neumann machine, so named after John von Neumann. Virtually all commercial computers are von Neumann machines. The functional architecture of such a machine includes a Central Processing Unit (CPU), an arithmetic unit, memory locations, and two kinds of memory access. The CPU has access to memory locations and current active data structures and determines what operations the computer is to perform by consulting instructions (programs) located in memory. A somewhat similar functional architecture is implicit in some symbolic models of mental abilities: online processing is done using a short-term memory store that holds information relevant to the process being carried out; and online processing influences and is influenced by long-term memory. Nevertheless, it is universally accepted by computationalists that the functional architecture of the mind is not von Neumann architecture.

Higher-level programming languages such as Basic, Pascal, FORTRAN, Cobol, Java, and C++, and Lisp (List Processing) are (or describe) universal machines that are automatic formal systems. Such languages are very useful since, as John Haugeland (1985) has aptly noted, some machines are easy to build and some are easy to program. So we build the machines that are easy to build and use them to simulate the machines that are easy to program. Higher-level programming languages are thus simulated machines relative to the machine language of the computers we actually build; they are thus virtual machines

relative to the machine language of the computer. (A complier is a program that translates the higher-level language into the machine language of the actual physical machine – in the case of von Neumann machines, into instructions encoded into strings of 1's and 0's.). No one thinks that any of these higher-level programming languages characterize the functional architecture of the mind.

Production Systems are Turing-machine equivalent higher-level programming languages used in research in the symbolic paradigm. Production Systems consist of rules specifying actions to be performed when certain conditions are met. While some computationalists think that at least some cognitive modules may have a Production System architecture, it is widely held by researchers in the symbolic paradigm that we do not yet know what the functional architecture of the mind is. The fundamental research objective of the symbolic paradigm is to determine what kind of automatic formal system the mind is.

Mind as syntactic engine

Recall that while a symbolic algorithm is defined over the formal properties of symbols (e.g., their shapes or syntactic structures), symbolic algorithms can govern symbol-to-symbol transitions that make sense given the meanings of the symbols. In proof theory, logical relations are treated as moves in a formal game; the formal moves preserve truth: they never take one from a truth to a falsehood. An algorithm can preserve truth: if one begins with a true symbol, the algorithm will take one only to a true symbol. The symbol-system paradigm is thus often said to presuppose a proof-theoretic conception of mind. Moreover, algorithms can be defined over the formal properties of symbols so as to preserve other sorts of semantic properties of the symbols, so that symbol transitions can implement reasoning processes of all sorts, not only deductive reasoning, but inductive reasoning, analogical reasoning, decision-making, etc. Haugeland (1985) suggests the following as the symbolist motto: "Take care of the syntax (the formal operations), and the semantics will take care of itself." As Daniel Dennett has aptly noted, the symbolic or rules and representations paradigm presupposes that the mind is "a syntactic engine."

The language of thought

Jerry Fodor and Zenon Pylyshyn have articulated a research program in the symbolic paradigm for reductively explaining propositional attitudes and mental processes involving them, a program that invokes the hypothesis that there is "a language of thought" (Fodor 1975; Pylyshyn 1986; Fodor & Pylyshyn 1988). Propositional attitudes have an intentional mode and a propositional content. The intentional modes include belief, desire, hope, wish, fear, intention, and the like. The propositional contents of propositional attitudes are expressed using that-clauses. Thus, the belief that that the cat is on the mat has the propositional content that the cat is one the mat; and the hope that it will not rain has the propositional content that it will not rain. According to the language-of-thought hypothesis, some mental symbols are atomic, and some are molecular in that they contain other mental symbols as constituents. Moreover, the mental symbol system has a compositional semantics: the content or meaning of a molecular symbol is a function of its syntactic structure and the contents of its constituent atomic symbols. The contents of propositional attitudes are expressed by sentence-like mental symbols; the contents of concepts, by word-like symbols. On this view, to be in a state within a certain intentional mode is to be disposed to compute in a certain way with an amalgam of concepts with a sentence-like syntactic structure – a mentalese sentence. Thus, to believe that *p* is to be disposed to compute in a certain way with a mentalese sentence that means that *p*; to desire that *p* is to be disposed to compute in a certain other way with a mentalese sentence that means that *p*, and so on for the other propositional attitudes. It is a topic for empirical investigation what the grammar of the language of thought is and what, exactly, these ways of being disposed to compute are.

As Fodor and Pylyshyn (1988) point out, if the mental symbol system includes a language of thought in the above sense, then we can appeal to it to explain, among other things, the systematicity and productivity of thought. They claim

that thought is productive in that one can, ideally, think an infinite number of thoughts; one is prevented from doing so only by the fact that one has a finite lifespan and finite memory resources. And they claim that thought is systematic in that pairs of abilities to have thoughts in the same intentional mode and with related contents are counterfactually dependent, in that one would have one member of the pair if one would have the other. Thus, normally, an individual that can think that Romeo loves Juliet can think that Juliet loves Romeo; and, normally, an individual that can hope the rock is next to the tree can hope the tree is next to the rock.

Psychosemantics

The question naturally arises of course as to how mental symbols – mental representations – acquire their contents or meanings. The numerals of various numeral systems such as Arabic numerals and Roman numerals derive their meanings from the linguistic conventions of communities. The current along the input and output wires of an electronic computer circuit derive their meanings from our intentional stipulative assignments; as do the keys of hand-calculators and the displays on their screen. If intentionality is to be explained by appeal to mental representations, then we need an account of their meaning that makes no appeal to linguistic conventions or semantic intentions, for these presuppose intentionality (Fodor 1990).

Procedural semantics and inferential or conceptual role semantics attempt to explain how mental representations derive their meaning from their participation in computational processes. John Searle's "Chinese Room" argument (see Chapter 9 above) makes a case that the semantic properties of symbols do not supervene on their syntactic relations, and so that intentional phenomena such as (e.g.) understanding Chinese cannot be explained solely by appeal to computational relations (see e.g. Searle 1999). If he is right about this failure of supervenience, then the kinds of semantics in question cannot succeed. Perhaps logical symbols – truth-functional symbols (e.g., "and" and "or") and symbols for the universal and existential quantifiers (respectively, "all" and "some") – can be defined by patterns of inferential relations. However, it seems that most symbols (e.g., "cow" and "run") cannot.

Semantic externalists claim that the meanings of many mental symbols fail to supervene on anything that goes on in our heads, since they depend on environmental relations. Externalists typically claim that a computational theory of intentionality must thus be supplemented with an externalist naturalistic account of meaning that makes no appeal to intentional notions. There are various programs for attempting to explain the semantic properties of mental symbols in purely naturalistic terms. However, the naturalistic relations appealed to (e.g., causal relations, counterfactual dependencies, information-theoretic relations, processes of natural selection), all appear to leave the semantic properties of mental representations underdetermined. Dual-aspect semanticists hope to combine inferential or conceptual role semantics with a naturalistic externalist account to determine the semantic properties of mental representations (Block 1986). This project, however, faces the problem of specifying which inferential relations are constitutive of the internal component of meaning (Fodor & LePore 1992). Moreover, it is uncertain whether any such combination will yield determinate meanings for mental symbols. Saul Kripke (1981) has argued that since our inferential dispositions are in fact finite (due to limitations in memory, our finite lifespan, etc.), they will leave the meanings of symbols radically indeterminate. Thus, suppose that one is in fact disposed to perform only 500 million calculations with a symbol "+". Even if the output of each of these 500 million calculations can be systematically interpreted as a computation of the plus function, one's dispositions to calculate will be compatible with the symbols expressing instead the "quus function," which can be defined as follows: x quus $y = x$ plus y, for any $y < 500{,}000{,}001$, otherwise x quus $y = 5$. Conceptual role theorists take this problem to be one of how conceptual roles should be idealized. But no entirely satisfactory solution to the Kripke's problem has yet been proposed. Suffice it to note that the problem of how mental representations acquire their meanings is perhaps the deepest problem that any reductive theory of intentionality faces.

3 The Connectionist Computational Paradigm

The functional architecture of our minds is somehow realized in our brains. One of the few things we know about the brain that seems to have a bearing on mentality is that it is, *inter alia*, a system of neural networks. Neural networks are not well understood; indeed neurons themselves are not well understood. There are somewhere between 50 and 500 different kinds of cortical neurons, and these different kinds of neurons appear to perform specialized functions not yet understood (Churchland & Sejnowski 1993). Nonetheless, just as top-down considerations of what cognitive functions are being computed can guide research into how mental abilities are seated in the brain, bottom-up considerations of the workings of neural networks can guide research into what algorithms are used to compute the functions in question. They can serve as a guide to discovering the functional architecture of the mind.

While the operations of actual neural networks are not well understood, the connectionist computational paradigm explores functional architectures that are at least "neurally inspired." Some connectionist networks, called "artificial neural networks," are specifically designed to approximate various kinds of real neural networks in various respects: the units (or nodes) in artificial neural networks are analogous to neurons, the connections among units analogous to axons, and the weights or strengths of the connections analogous to synapses. These networks are extensively employed in the growing field of computational neuroscience (see Churchland & Sejnowski 1993). The connectionist networks employed to model the computation of cognitive functions typically make no attempt at neural realism. Nonetheless, they have at least "a neural flavor."

Connectionist networks

A connectionist network is composed of interconnected units (or nodes). Individual units do all the information processing: there is thus no executive or CPU. There is, moreover, no program stored in memory; the program is implicit in the pattern of connectivity exhibited by the units. Many units process information simultaneously, and so the network as a whole engages in parallel distributed processing (PDP) (see McClelland & Rumelhart et al. 1986; Rumelhart & McClelland et al. 1986).

Units have states of activation. Depending on the network architecture, a unit may have only two states of activation, "on" and "off," three or more discrete states of activation, or continuous levels of activation, bounded or unbounded. Units process information by changing or retaining their state of activation in response to activation signals from their sending units.

The connections among units can be of various strengths or weights, which affect the extent to which the activity output of a sending unit influences the activation states of receiving units. Connections can, moreover, be excitatory or inhibitory. The connection from a unit U_i to a unit U_j is excitatory if the activation output of U_i tends to increase the level of activation of U_j, and inhibitory if it tends to decrease it. The causal influence exerted by a sending unit on a receiving unit thus depends on two intrinsic properties of their connection: its weight and whether it is excitatory or inhibitory. The notation "w_{ij}" is used to stand for a real number that indexes the connection from unit U_i to unit U_j by its weight and kind. The number is positive when the connection is excitatory, negative when it is inhibitory. The weight of the connection is the absolute value of w_{ij}. In many networks, the extent and kind of causal influence a unit U_i exerts on a unit U_j is indexed by the product of w_{ij} and the activation value of U_i.

Whether a unit changes or retains its state of activation is a function of three factors: the totality of input to the unit, the unit's current activation state, and the unit's bias (if any), which may be positive or negative. A unit thus computes an activation function that maps its total network activity input (and external input if any), its current activation state, and its bias (if any) to an activation state. The total network activity input to U_j is the sum of all of the causal influences from sending units: it is the sum of the product of each activation value from a sending unit and the real number indexing the weight

and kind of connection the sending unit bears to U_i. While the output function of a unit can be linear, it is typically a threshold function, and thus nonlinear. If a unit has a negative bias, it may send 0, the null signal, unless its activation value exceeds a certain threshold. If it has a positive bias, it may send a certain non-0 activation value unless its activation state value falls below a certain threshold. In Hopfield networks, units have a sigmoidal (S-shaped) response to net input: their output only increases by a given amount given an increase in net input; after that, they increase no further. But units can also have Gaussian (bell-shaped) output functions and other sorts of nonlinear ones as well.

The input units of a network receive signals directly from the environment of the network, while the output units send signals directly to the environment. Since input and output units directly interact with the environment, they are called "visible units." So-called "hidden units" only directly interact with other units.

Some networks, called "perceptrons," have only two layers of units: a layer of input units and a layer of output units. Minsky and Pappert (1969) showed perceptrons are limited in their computational power: for example, they cannot compute XOR (exclusive-or). The reason is that the problem of determining the value of the XOR function is a linearly inseparable problem and perceptrons cannot solve any such problem. Rumelhart et al. 1986 demonstrated, however, that networks with three or more layers can compute XOR and, more generally, can solve decidable linearly inseparable problems.

Feedforward networks with one or more layers of hidden units are called "multilayered" networks. The Hamming net, for instance, is a widely used feedforward network with three layers of units, one of which is a layer of hidden units. Feedforward networks are non-interactive: activation flows from the input units through each layer of hidden units to the output units. In interactive networks, two units can mutually influence each other, and thus a unit can be related to another both as an input unit and as an output unit. In interactive networks, two-way connections between units are often symmetrical, so that $w_{ij} = w_{ji}$. In competitive networks, units form pools: the units in a pool are all mutually inhibitory, while units outside of the pool bear excitatory connections to one or more units in the pool. In recurrent networks, there are connection patterns that contain loops, so that a unit is either related to itself as an input unit or there is a series of connections from the unit back to itself, so that the output of a unit at one time can causally influence its activation state at another. In auto-associative networks, each unit is connected to every other unit, including itself. These are only some of the many kinds of patterns of connectivity that are possible.

The behavior of a network as a whole is a consequence of the pattern of connectivity exhibited by its units at a time and the global activation state of the network at that time. A vector (ordered list) of real numbers is used to index the global activation state. The activation value of each unit in the network at that time is indexed by one and only one element of the vector, and so the number of elements in the vector will be equal to the number of units in the network. The set of all possible global activation states of a network is its activation space, whose dimensionality is exactly equal to the number of units in the network. A network with n units will thus be indexed by an n-tuple of real numbers that identifies a position in an n-dimensional vector space. That position represents the global activation state of the network at a time. A temporal series of global activation states will trace a path through a vector space. Network information processing is characterized as the evolution through time of global patterns of activation. Networks can be systematically interpreted as performing mathematical operations on matrices such as matrix multiplication (e.g., a computing inner product).

Explicit representations in a connectionist network can be either local or distributed. Local representations are individual units, or individual units at certain levels of activation. Distributed representations are patterns of activity over a group of units. The pattern of connectivity of a network is sometimes characterized as implicitly representing.

As a result of the pattern of connectivity exhibited by its units, a network as a whole can behave in rule-like ways to compute functions. When a feedforward network computes a

function, the arguments of the function are represented by different patterns of activity over the input units, and the values of those arguments for the function by patterns of activity over output units. In computing the function, the network acts as a look-up table. Input activation patterns function like questions posed to the network (for example, "To what category does this belong?"), and output patterns function like answers to the question ("To category C"). Unlike in a symbolic look-up table, however, the answers are not stored as data structures; rather they are implicitly represented in the pattern of connectivity. (Sometimes, however, hidden units are explicit representations.) In interactive units, the argument of a function might be represented by an initial pattern of activation over units in the network, and the value of the function for that argument by a pattern of activation over those same units, a pattern the network "settles" or "relaxes" into after information processing.

The "neural flavor" of connectionist networks is by no means their only attraction for computationalists. Connectionist networks are good at pattern recognition tasks: pattern matching, pattern completion, and pattern association. They degrade gracefully in their performance of a task as a network is damaged (e.g., when a unit is lost) and in the face of noisy or incomplete data. They can learn. That is to say, with proper training, they can improve their performance of various tasks. Moreover, they are very efficient at solving connected problems and at arriving at optimal or near optimal solutions to best-match problems. Of these attractions, more below.

Learning

The weights of the connections in a network are not architecturally fixed. Learning consists in changes in the weights. There are various learning rules that govern weight change. The Hebb Rule is based on Donald Hebb's (1949) hypothesis that the connections between two neurons might strengthen whenever they fire simultaneously. According to the Hebb rule, the weight of a connection between units should be increased or decreased in proportion to the products of their simultaneous activations (Rumelhart et al. 1986). The Delta rule is an error correction rule that changes the weights leading from units sending signals to output units on the basis of the discrepancy between the actual output and the desired one. Backpropagation is a generalization of the Delta rule, and is widely used in multilayered networks. The actual output activation pattern for a given input activation pattern is compared with the desired output. The difference between the two is then propagated back into whatever connections were used to get the actual output activation pattern. The connections among units that contributed to the actual output are strengthened (increased in weight) when the match is good, and are reduced in strength (decreased in weight) when it is poor. The weights are thus adjusted so as to reduce the margin of error between the actual output and the desired one. And in this way, the network learns.

Network training can be supervised or unsupervised. In supervised learning, the network is provided explicit feedback from an external source about what output is desired as a response to a certain input. The Delta rule and backpropagation are used in supervised learning. In unsupervised learning, no such external feedback is provided to the network; rather, the network monitors its own performance through internal feedback. There are a number of unsupervised learning algorithms; the Kohonen algorithm is one such (see Beale & Jackson 1990, ch. 5).

Pattern recognition

NETtalk offers a vivid example of the ability of networks to learn to recognize patterns. Designed by Terrence Sejnowski and C. R. Rosenberg (1987), it is a network that learns to map letters onto phonemes. NETtalk is a three-layered feedforward network, whose input units represent letters (individual letters are represented by patterns of activation over 29 input units and there are 7 such groups of 29 input units) and whose output units represent phonemes. The network feeds into a synthesizer. After sufficient training using backpropagation, when presented strings of letters comprising the words of actual

English text, the network drives the synthesizer to sound like a robotic voice literally reading the text.

Some studies have compared networks using backpropagation with various members of the class of "top-down inductive decision trees" ("decision trees" for short) in respect of accuracy in pattern recognition and the ability to deal with noisy, incomplete data gracefully, as well as in other respects (see Sethi & Jain 1991). Decision trees are production-rule systems, and thus symbolic systems, that excel at pattern recognition tasks (see Utgoff 1999). One comparative study by Shavlik et al. (1991) included a comparison between a decision tree system called "ID3" and a network using backpropagation on both the NETtalk A data set and the NETtalk full data set. Shavlik et al. noted: "for the most part [both] learning systems are similar with respect to accurately classifying novel examples" (1991: 120). They also compared ID3 and the network on three kinds of noisy data: inaccurate feature values, missing feature values, and insufficient number of features. The network performed slightly better than ID3 on inaccurate feature values and missing feature values, but when trained on insufficient numbers of feature values, "ID3 and backpropagation degrade at roughly the same rate as the number of features is reduced" (1991: 130). Several other comparative studies have found that networks using backpropagation are roughly comparable to various other members of the family of decision trees as concerns accuracy in pattern recognition and graceful degradation in response to noisy, incomplete data. Decisions trees are, however, much faster at learning than networks using backpropagation (see Sethi & Jain 1991; Marinov 1993; and McLaughlin & Warfield 1994).

Connected problems, best-match problems, and multiple soft constraint satisfaction

Networks are especially efficient at solving connected problems – problems that do not divide into independently solvable subproblems, like the traveling salesman problem. The goal is to find the shortest route that a salesman can take to visit each of a number of cities, while visiting each city only once. Since which city the salesman visits depends on which cities he has already visited, the problem does not divide into independently solvable subproblems. The traveling salesman problem is decidable, and so can be solved by a symbol system. But if, for instance, there are 20 cities to visit, there is a minimum of 653,837,184,000 possible routes to take (Raggett & Bains 1992). Thus, as the number of cities increases there is an exponential increase in the computational resources required to solve the traveling salesman problem within a symbol system (see Chapter 2, Complexity). A Hopfield network is able to find a solution in a small number of training cycles.

Networks are also especially efficient at finding optimal or near optimal solutions to what Minsky and Papert (1969) called "best-match problems" – problems whose solution involves assessing the satisfaction of multiple soft (i.e., non-mandatory) constraints. Hinton (1977) developed the first network approach to solving best-match problems. The following description of it paraphrases the description provided in McClelland and Rumelhart 1988. Units in the network have one of two states of activation ("on" or "off"), and the network contains local representations, each unit representing a different hypothesis. The connections in the network implicitly represent evidential relationships among the hypotheses. Thus if a hypothesis H is evidence for another hypothesis H', the connection from the unit representing H to the unit representing H' is positive; if H is evidence against H', the connection is negative. The stronger the confirming or disconfirming relationship the one hypothesis bears to the other, the stronger the connection between the units that represent them. If one hypothesis H entails another H', then the connection between the unit representing H and the unit representing H' will be such that the second unit is on whenever the first is on. When two hypotheses are incompatible, they are connected in such a way that they cannot both be on. Since the network contains input units that receive signals from the environment, hypotheses can receive confirming or disconfirming evidence directly from the environment. The fact that different hypotheses have different *a priori*

probabilities is captured by biasing the relevant units. The overall goodness of fit of a particular hypothesis to evidence is measured by the sum of the individual constraints on the activation value of the unit representing the hypothesis. The activation values of units range between a minimum and a maximum. The maximum value means that the hypothesis should be accepted, the minimum that it should be rejected, and intermediate values correspond to intermediate levels of certainty. The constraint satisfaction problem is thus reduced to one of maximizing this overall goodness of constraint fit.

Of course, this approach to best-match problems may well not seem interestingly different from a symbolic approach; indeed it might be viewed as a way of implementing a symbolic decision procedure on a network. The computationally relevant formal property of a symbol can be a unit's being activated, and symbols can participate in probabilistic algorithmic processes. There are other networks, however, that solve best-match problems in strikingly different, and striking efficient ways.

A problem that arises for procedures for solving best-match problems is that of avoiding local maxima of goodness of constraint fit. It can be characterized as an energy minimization problem (McClelland & Rumelhart 1988). The analog of the goodness maximum is the energy minimum, and the analogs of local goodness maxima are local energy minima. The situation is easy to visualize as an energy landscape. In an energy landscape, the goodness maximum corresponds to the lowest valley in the landscape, while local goodness minima correspond to local valleys in the landscape. The problem of avoiding local goodness maxima is thus the problem of avoiding settling into a local valley, rather than into the lowest valley in the energy landscape. John Hopfield (1982) observed that some networks behave in such a way as to minimize a measure over the whole network, one he aptly called "energy." The behavior of these networks can be characterized as moving through an energy landscape. The Boltzmann machine, developed by Hinton and Sejnowski (1986) and named after the physicist Ludwig Boltzmann, is such a network. To handle the problem of local maxima of goodness of constraint fit, the Boltzmann machine employs a computational analog of the metallurgical process of annealing. Annealing is a process whereby metals are heated to a little below their melting point and then cooled very slowly so that all their atoms have time to settle into a single orientation. The analog of heat in the Boltzmann machine is random noise that is introduced into network activity. The function of the noise is to jar the network out of local valleys, so that it can explore other parts of the energy landscape to find the lowest valley, thereby achieving the global maximum of fit. When the network reaches a stable state, it has settled or relaxed into a solution. Given sufficient time, the Boltzmann machine can find the energy minimum for any best-match problem.

Networks and cognitive abilities

There are network models of certain aspects of motor control and low-level perception. For example, there is a network model for the vestibular ocular reflex, which enables eyes to track an object as the head moves (Churchland & Sejnowski 1993). As concerns low-level visual perception, the linear models color-constancy algorithm, for instance, is characterized in connectionist terms (Maloney & Wandell 1986). Moreover, opponent processing theory, the leading neuro-computational theory of color experience, is easily understood in connectionist terms (Hurvich 1981). According to the theory, there are pairs of opponent channels that respond differently to the outputs of our three types of cones (L-cones, M-cones, and S-cones): the RED channel and the GREEN channel, the BLUE channel and the YELLOW channel, and the BLACK channel and the WHITE channel. Activity in one channel inhibits activity in its opponent channel. The theory explains the fact that nothing looks bluish-yellow as a result of the fact that activity in the BLUE channel inhibits activity in the YELLOW channel and conversely. And the theory explains the appearance of unique blue (blue that is not at all reddish or greenish) as the result of the activation of the BLUE channel, the YELLOW channel being deactivated, and the RED and GREEN channels being in equilibrium. These channels can, of course, be

understood as pools of interconnected neuron-like units.

Connectionist modeling has by no means been restricted to motor control and low-level perception. It has been extended to most areas of cognition. Connectionist networks have been appealed to as mechanisms for processes of categorization in terms of resemblance to prototypes (Churchland 1995). And there are, moreover, connectionist models of various aspects of language comprehension and production. For example, Jeffrey Elman (1990) has explored how recurrent nets can learn to become sensitive to anaphoric relations. And Alan Prince and Paul Smolensky's (1993) optimality theory, for instance, outlines how a multiple soft constraint satisfaction might serve as a natural language parser might work by employing multiple soft constraint satisfaction. This sample represents only a tiny fraction of connectionist cognitive modeling. (For a recent very brief overview, see McClelland 1999.)

4 How are Paradigms Related?

The relationship between the symbolic and connectionist paradigms is a topic of heated dispute. It will have to suffice here simply to list some possibilities.

One possibility is that the functional architecture of the mind is hybrid: some aspects of cognition are symbolic and some are connectionist. For example, perhaps motor control modules and low-level perceptual modules have a connectionist architecture, while linguistic modules and reasoning in central processing have a symbolic architecture. On this view, the mind is not a single kind of computer, but has different kinds of compartments that are different kinds of computers.

Another possibility is that the functional architecture of the mind is a symbolic architecture, but this architecture is implemented by a connectionist one. (Both Turing machines and Production Systems, it should be mentioned, have been implemented by connectionist networks.) Connectionism, it is often claimed, is concerned with microcognition; and its advocates often characterize the connectionist paradigm as the subsymbolic paradigm. Some connectionists have suggested that the relationship between the symbolic paradigm and the connectionist (subsymbolic) paradigm is analogous to the relationship that Newtonian physics bears to quantum mechanics (Rumelhart et al. 1986). But proponents of the implementational view of connectionism will claim that the more apt analogy is the relationship chemistry bears to quantum mechanics. Atoms are constituted by subatomic particles, and chemical processes are implemented by quantum mechanical ones (indeed all causal processes are ultimately implemented by quantum mechanical ones). Perhaps, analogously, atomic symbols are constituted by patterns of activation, and connectionist processes implement symbolic algorithmic processes (McLaughlin 1993a).

Yet another possibility is that the functional architecture of the mind is either thoroughly symbolic or thoroughly connectionist. Proponents of this view see the two paradigms as in a sort of zero-sum competition. The leading objection to the view that the functional architecture of the mind is thoroughly connectionist is due to Fodor and Pylyshyn (1988). They argue that a connectionist architecture cannot explain the systematicity of thought without implementing a symbolic architecture. A large literature has arisen in response to this objection. Some deny that thought is systematic. Some concede that thought is systematic, but maintain that it is not the job of a theory of cognitive architecture to explain systematicity (see McLaughlin 1993b for a discussion). And some attempt to show how a connectionist architecture could explain the systematicity of thought without implementing a symbolic one (see e.g. Smolensky 1991; for a response, see Fodor & McLaughlin 1990; McLaughlin 1997a; for a counter-response, see Cummins 2000).

Another possibility still is that the functional architecture of the mind somehow integrates features of symbolic architectures and features of connectionist architectures. Smolensky (1994, 1995) has sketched an architecture he calls "an Integrated Connectionist Symbol architecture (ICS)" that attempts to include features of both connectionist and symbol architectures, and which he claims will explain the systematicity and

productivity of thought. But it remains a question whether when developed sufficiently to explain systematicity and productivity, it would collapse into an implementation architecture for a symbolic one (McLaughlin 1997a).

Of course, not all of these possibilities are exclusive. It may be that some cognitive modules are thoroughly connectionist, and that various other modules are symbolic but implemented by connectionist networks. There is, it should be noted, a growing body of work in the field of machine learning that attempts to develop machines with a mixed architecture, including symbolic and connectionist subcomponents. Of course, this work in machine learning is not concerned with either psychological or neurological realism (but rather with building better machines). Nonetheless, it demonstrates the usefulness of such mixed architectures for computational purposes.

Suffice it to note that the functional architecture of the mind remains an open question. What the future will bring remains to be seen.

References

Beale, R. and Jackson, T. 1990. *Neural Computing*. Bristol: Adam Hilger. [An excellent introduction to neural network computing that is accessible to advanced undergraduates.]

Bechtel, W. and Abrahamsen, A. 1991. *Connectionism and the Mind*. Oxford: Blackwell. [An excellent introduction to connectionist networks and their contribution to the understanding of mentality, and a major source of the discussion of the connectionist paradigm in this chapter. Accessible to advanced undergraduates.]

Block, N. 1981. "Psychologism and behaviorism." *Philosophical Review* 90: 5–43. [A defense of the view that mentality does not supervene on any pattern of dispositions to outward or peripheral behavior. Accessible to advanced undergraduates.]

——. 1986. "An advertisement for a semantics for psychology." *Midwest Studies in Philosophy* 10: 615–78. [A seminal defense of conceptual role semantics. Accessible to advanced undergraduates.]

Chalmers, D., ed. 1996. *Psyche*, vol. 2, <http://Psyche.cs.monash.edu.au/>. [Contains responses to Penrose 1989. Accessible to graduate students and professionals.]

Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey Into the Brain*. Cambridge, MA: MIT/Bradford. [Discussion of the role of networks in understanding mentality. Accessible to advance undergraduates.]

Churchland, P. S. and Sejnowski, T. J. 1993. *The Computational Brain*. Cambridge, MA: Bradford/MIT. [An introduction to neural network modeling in the field of computational neurscience. Accessible to advanced undergraduates.]

Copeland, J. B. 2000. "Narrow versus wide mechanism: including a re-examination of Turing's views on the mind–machine issue." *Journal of Philosophy* 1: 5–32. [A defense of the view that a mechanistic view of the mind need not treat minds as Turing equivalent computational mechanisms. Accessible to advanced undergraduates.]

Cummins, R. and Schwarz, G. 1991. "Connectionism, computation, and cognition." In Horgan & Tienson 1991: 60–73. [A discussion of the role of connectionism in the computational theory of mind. Accessible to advanced undergraduates.]

Elman, J. L. 1990. "Finding structure in time." *Cognitive Science* 14: 179–211. [An attempt to show how connectionist networks can learn to compute certain linguistic functions. Accessible to graduate students.]

Fodor, J. 1975. *The Language of Thought*. New York: Crowell. [A seminal defense of the language-of-thought hypothesis. Accessible to advanced undergraduates.]

——. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT/Bradford. [A seminal discussion of issues in psychosemantics. Accessible to advanced undergraduates.]

—— and LePore, E. 1992. *Holism*. Malden, MA: Blackwell. [Contains criticisms of inferential/conceptual role semantics. Accessible to advanced undergraduates.]

—— and McLaughlin, B. P. 1990. "Connectionism and the problem of systematicity: why Smolensky's solution doesn't work." *Cognition* 35: 183–204. Repr. in Horgan & Tienson 1991. [A response to Smolensky 1991. Accessible to advanced undergraduates.]

—— and Pylyshyn, Z. 1988. "Connectionism and cognitive architecture." *Cognition* 28: 183–204. [A classical attack on the view that cognitive architecture may be thoroughly connectionist. Accessible to advanced undergraduates.]

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT/Bradford.

[An excellent introduction to the symbolic approach to artificial intelligence. Accessible to undergraduates.]

——, ed. 1997. *Mind Design II*. Cambridge, MA: MIT/Bradford. [Contains classical articles in the field of artificial intelligence and the computational theory of mind. Accessible to advanced undergraduates.]

Hebb, D. 1949. *The Organization of Behavior*. New York: Wiley & Sons. [A classical work on neural networks. Accessible to advanced undergraduates.]

Hinton, G. E. and Sejnowski, T. J. 1986. "Learning and relearning in Boltzmann machines." In Rumelhart et al. 1986: 282–317. [A presentation of Boltzmann machines. Accessible to graduate students.]

Hopfield, J. J. 1982. "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the National Academy of Sciences, USA*, 79: 2554–8. [Presents what have come to be called "Hopfield nets." Accessible to graduate students.]

Horgan, T. and Tienson, J., eds. 1991. *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer. [A collection of essays on philosophical issues concerning connectionism. Accessible to advanced undergraduates.]

Hurvich, L. M. 1981. *Color Vision*. Sunderland, MA: Sinauer.

Jackson, J. 1993. "Appendix A [for philosophers]." *Philosophy and Phenomenological Research* 53: 901–6. [Argues that mentality does not supervene on dispositions to outward or peripheral behavior. Accessible to advanced undergraduates.]

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press. [Contains a presentation of the plus–quus problem for computationalists. Accessible to advanced undergraduates.]

Lewis, D. 1979. "Lucas against mechanism II." *Canadian Journal of Philosophy* 9: 373–6. [Response to Lucas 1961. Accessible to graduate students and professionals.]

Lucas, J. R. 1961. "Minds, machines and Gödel." *Journal of Philosophy* 36: 112–27. [Appeals to the results of Kurt Gödel to argue that we are not computationally equivalent to Turing machines. Accessible to graduate students.]

Macdonald, C. and Macdonald, G., eds. 1995. *Connectionism: Debates on Psychological Explanation*. Malden, MA: Blackwell. [A collection of articles on philosophical issues surrounding the connectionist approach to the mind. Accessible to advanced undergraduates.]

Maloney, L. T. and Wandell, B. A. 1986. "Color constancy: a method for recovering surface spectral reflectance." *Journal of the Optical Society of America* A 3: 29–33. [Presents the linear models color-constancy algorithm. Accessible to graduate students.]

Marinov, M. S. 1993. "On the spuriousness of the symbolic/subsymbolic distinction." *Minds and Machines* 3: 253–70. [Contains a discussion of comparative studies of connectionist networks and decision trees. Accessible to advanced undergraduates.]

Marr, D. 1982. *Vision*. New York: Freeman. [A seminal work on low-level vision. Accessible to graduate students.]

McCall, S. 1999. "Can a Turing machine know that the Gödel sentence is true?" *Journal of Philosophy* 10: 525–32. [Appeals to results of Kurt Gödel to argue that we are not computationally equivalent to Turing machines. Accessible to graduate students.]

McClelland, J. L. 1999. "Cognitive modeling, connectionist." In Wilson & Kiel 1999: 137–41. [A brief overview of connectionist modeling. Accessible to advanced undergraduates.]

—— and Rumelhart, D. E. 1988. *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. Cambridge, MA: MIT/Bradford. [An introduction to connectionist modeling, and a major source of the discussion of the connectionist paradigm in this chapter. Accessible to advanced undergraduates.]

——, ——, and the PDP Group, eds. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 2, *Psychological and Biological Models*. Cambridge, MA: MIT/Bradford. [One of the two "bibles" of connectionism. This seminal collection of essays that introduces and develops the connectionist paradigm. Some of the chapters are accessible to advanced undergraduates, but some are more appropriate for graduate students and professionals.]

McLaughlin, B. P. 1993a. "The connectionism/classicism battle to win souls." *Philosophical Studies* 70: 45–72. [An examination of the debate between the symbolic paradigm and the connectionist paradigm. Accessible to advanced undergraduates.]

——. 1993b. "Systematicity, conceptual truth, and evolution." In C. Hookway and D. Peterson, eds., *Royal Institute of Philosophy, Supplement* 34:

217–34. [Examines some attempts to show that it is not the job of the computational theory of mind to explain the systematicity of thought. Accessible to advanced undergraduates.]

——. 1997a. "Classical constituents in Smolensky's ICS architecture." In M. L. dalla Chiara, K. Doetes, D. Mundici, and J. van Bentham, eds., *The Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995*, vol. 2, *Structures in Science*. Dordrecht: Kluwer, 331043. [Argues that Smolensky's ICS architecture can containmental representations with classical constituents. Accessible to advanced undergraduates.]

——. 1997b. "Connectionism." In E. Craig, ed., *The Routledge Encyclopedia of Philosophy*. London: Routledge. [An introduction to the connectionist paradigm. A source for the discussion of the connectionist paradigm in this chapter.]

—— and O'Leary-Hawthorne, J. 1994. "Dennett's logical behaviorism." *Philosophical Topics* 22: 189–258. [Presents Daniel Dennett's intentional systems theory, and argues that, *contra* that theory, intentionality does not supervene on any pattern of dispositions to outward or peripheral behavior. Accessible to advanced undergraduates.]

—— and Warfield, T. A. 1994. "The allure of connectionism reexamined." *Synthese* 101: 365–400. [Contains a discussion of studies comparing connectionist networks and decision trees. Accessible to advanced undergraduates.]

Minsky, M. A. and Pappert, S. 1969. *Perceptrons*. Cambridge, MA: MIT Press. [A critical examination of perceptrons, two-layered connectionist networks. Accessible to graduate students.]

Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press. [Includes an appeal to results of Kurt Gödel to argue that our minds are not computationally equivalent to Turing machines. Accessible to graduate students.]

Prince, A. and Smolensky, P. 1993. *Optimality Theory: Constraint and Interaction in Generative Grammar*. Technical report TR-2, New Brunswick, NJ: Rutgers Center for Cognitive Sciences. [A seminal presentation of optimality theory. Accessible to graduate students.]

Putnam, H. 1975. "Minds and machines." In his *Philosophical Papers*, vol. 2, *Mind, Language, and Reality*. Cambridge: Cambridge University Press. [A classical paper on machine state psychofunctionalism. Accessible to advanced undergraduates.]

Pylyshyn, Z. 1986. *Computation and Cognition*. Cambridge, MA: MIT/Bradford. [A seminal discussion of the symbolic paradigm, and a major source for the discussion of that paradigm in this chapter. Accessible to advanced undergraduates.]

Raggett, J. and Bains, W. 1992. *Artificial Intelligence From A to Z*. London: Chapman & Hall. [A wonderful source of very simple explanations of technical terms in the field of artificial intelligence. Accessible to advanced undergraduates.]

Ramsey, W., Stich, S., and Rumelhart, D., eds. 1991. *Philosophy and Connectionism Theory*. Hillsdale, NJ: Erlbaum. [A collection of articles on the connectionist paradigm. Accessible to advanced undergraduates.]

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, eds. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, *Foundations*. Cambridge, MA: MIT/Bradford. [One of the two "bibles" of connectionism. A classical introduction of the connectionist paradigm, and a major source for the discussion of that paradigm in this chapter. Some chapters are accessible to advanced undergraduates, but some are most appropriate for graduate students and professionals.]

Searle, J. 1999. "Chinese room argument." In Wilson & Keil 1999: 115–16. [Searle's most recent presentation of his well-known Chinese room argument. Accessible to undergraduates.]

Sejnowski, J. and Rosenberg, C. R. 1987. "Parallel networks that learn to pronounce English text." *Complex Systems* 1: 145–68. [Introduces NETtalk. Accessible to graduate students.]

Sethi, I. K. and Jain, A. K., eds. 1991. *Artificial Neural Networks and Statistical Pattern Recognition*. Amsterdam: Elsevier/North Holland. [A collection of essays including some comparative studies of connectionist networks and decision trees. Accessible to graduate students.]

Shavlik, J. W., Mooney, R. J., and Towell, G. G. 1991. "Symbolic and neural learning algorithms: an experimental comparison." *Machine Learning* 6: 111–44.

Smolensky, P. 1991. "The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn." In Horgan & Tienson 1991: 281–308. [A reply to Fodor & Pylyshyn 1988. Accessible to advanced undergraduates.]

——. 1994. "Computational models of mind." In Guttenplan 1994: 176–84. [An introduction to the connectionist paradigm that includes a sketch

of Smolensky's ICS architecture. Accessible to advanced undergraduates.]

——. 1995. "Reply: constitutive structure and explanation in an integrated connectionist/symbolic cognitive architecture." In Macdonald & Macdonald 1995: 223–90.

Sterelny, K. 1990. *The Representational Theory of Mind: An Introduction*. Oxford: Blackwell. [An excellent introduction to the representational theory of mind. Accessible to advanced undergraduates.]

Turing, A. 1950. "Computing machinery and intelligence." *Mind* 59: 433–60. [A truly groundbreaking article that literally began the field of artificial intelligence. Accessible to undergraduates.]

Utgoff, P. 1999. "Decision trees." In Wilson & Kiel 1999: 223–5. [An introduction to decision trees. Accessible to advanced undergraduates.]

Van Gelder, T. 1995. "What might cognition be, if not computation?" *Journal of Philosophy* 7: 345–81. [A defense of the view that cognition should be understood in dynamical systems terms, rather than computationally. Accessible to advanced undergraduates.]

Wilson, R. A. and Keil, F. C., eds. 1999. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT/Bradford. [Contains many entries relevant to topics discussed in this article. Entries are accessible to advanced undergraduates.]

Part IV

Real and Virtual Worlds

Chapter 11

Ontology

Barry Smith

Philosophical Ontology

Ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes, and relations in every area of reality. "Ontology" is often used by philosophers as a synonym of "metaphysics" (a label meaning literally: "what comes after the *Physics*"), a term used by early students of Aristotle to refer to what Aristotle himself called "first philosophy." Sometimes "ontology" is used in a broader sense, to refer to the study of what *might* exist; "metaphysics" is then used for the study of which of the various alternative possible ontologies is in fact true of reality (Ingarden 1964). The term "ontology" (or *ontologia*) was coined in 1613, independently, by two philosophers, Rudolf Göckel (Goclenius) in his *Lexicon philosophicum* and Jacob Lorhard (Lorhardus) in his *Theatrum philosophicum*. Its first occurrence in English as recorded by the *Oxford English Dictionary* appears in Bailey's dictionary of 1721, which defines ontology as "an Account of being in the Abstract."

Ontology seeks to provide a definitive and exhaustive classification of entities in all spheres of being. The classification should be definitive in the sense that it can serve as an answer to such questions as: What classes of entities are needed for a complete description and explanation of all the goings-on in the universe? Or: What classes of entities are needed to give an account of what makes true all truths? It should be exhaustive in the sense that all types of entities should be included in the classification, including also the types of relations by which entities are tied together to form larger wholes.

Different schools of philosophy offer different approaches to the provision of such classifications. One large division is that between what we might call substantialists and fluxists, which is to say between those who conceive ontology as a substance- or thing- (or continuant-) based discipline and those who favor an ontology centered on events or processes (or occurrents). Another large division is between what we might call adequatists and reductionists. Adequatists seek a taxonomy of the entities in reality at all levels of aggregation, from the microphysical to the cosmological, and including also the middle world (the *mesocosmos*) of human-scale entities in between. Reductionists see reality in terms of some one privileged level of existents; they seek to establish the "ultimate furniture of the universe" by decomposing reality into its simplest constituents, or they seek to "reduce" in some other way the apparent variety of types of entities existing in reality.

It is the work of adequatist philosophical ontologists such as Aristotle, Ingarden (1964), and Chisholm (1996) which will be of primary importance for us here. Their taxonomies are in many ways comparable to the taxonomies produced by sciences such as biology or chemistry, though they are of course radically more general than these. Adequatists transcend the dichotomy between substantialism and fluxism, since they accept categories of both continuants and occurrents. They study the totality of those objects, properties, processes, and relations that make up the world on different levels of focus and granularity, and whose different parts and moments are studied by the different scientific disciplines. Ontology, for the adequatist, is then a descriptive enterprise. It is thus distinguished from the special sciences not only in its radical generality but also in its goal or focus: it seeks not predication and explanation but rather taxonomy and description.

Methods of Ontology

The methods of ontology – henceforth in philosophical contexts always used in the adequatist sense – are the methods of philosophy in general. They include the development of theories of wider or narrower scope and the testing and refinement of such theories by measuring them up, either against difficult counter-examples or against the results of science. These methods were familiar already to Aristotle himself.

In the course of the twentieth century a range of new formal tools became available to ontologists for the development and testing of their theories. Ontologists nowadays have a choice of formal frameworks (deriving from algebra, category theory, mereology, set theory, topology) in terms of which their theories can be formulated. These new formal tools, along with the language of formal logic, can in principle allow philosophers to express intuitive ideas and definitions in clear and rigorous fashion, and, through the application of the methods of formal semantics, they can allow also for the testing of theories for logical consistency and completeness.

Ontological Commitment

To create effective representations it is an advantage if one knows something about the things and processes one is trying to represent. (We might call this *the Ontologist's Credo*.) The attempt to satisfy this credo has led philosophers to be maximally opportunistic in the sources they have drawn upon in their ontological explorations of reality and in their ontological theorizing. These have ranged all the way from the preparation of commentaries on ancient texts to reflection on our linguistic usages when talking about entities in domains of different types. Increasingly, however, philosophers have turned to science, embracing the assumption that one (perhaps the only) generally reliable way to find out something about the things and processes within a given domain is to see what scientists say. Some philosophers have indeed thought that the way to do ontology is exclusively through the investigation of scientific theories.

With the work of Quine (1953) there arose in this connection a new conception of the proper method of ontology, according to which the ontologist's task is to establish what kinds of entities scientists are committed to in their theorizing. The ontologist studies the world by drawing conclusions from the theories of the natural sciences, which Quine takes to be our best sources of knowledge as to what the world is like. Such theories are extensions of the theories we develop and use informally in everyday life, but they are developed with closer attention to those special kinds of evidence that confer a higher degree of probability on the claims made. Quine – or at least the Quine of 1953 (I am here leaving aside Quine's views on such matters as ontological relativity and the indeterminacy of translation) – still takes ontology seriously. His aim is to use science for ontological purposes, which means: to find the ontology *in* scientific theories. Ontology is then a network of claims, derived from the natural sciences, about what exists, coupled with the attempt to establish what types of entities are most basic. Each natural science has, Quine holds, its own preferred repertoire of types of objects to the existence of which it is committed. Each such science

embodies only a partial ontology. This is defined by the vocabulary of the corresponding theory and (most importantly for Quine) by its canonical formalization in the language of first-order logic. Note that ontology is for Quine himself not the metalevel study of the ontological commitments or presuppositions embodied in the different natural-scientific theories. Ontology is rather these commitments themselves. Quine moves to the metalevel, making a semantic ascent to consider the statements in a theory, only in setting out to establish those expressions which definitively carry its commitments. Quine fixes upon the language of first-order logic as the medium of canonical representation not out of dogmatic devotion to this particular form, but rather because he holds that this is the only really clear form of language. First-order logic is itself just a regimentation of corresponding parts of ordinary language, a regimentation from which, in Quine's eyes, logically problematic features have been excised. It is then, Quine argues, only the bound variables of a theory that carry its definitive commitment to existence. It is sentences like "There are horses," "There are numbers," "There are electrons," that do this job. His so-called "criterion of ontological commitment" is captured in the slogan: *To be is to be the value of a bound variable*. This should not be understood as signifying some reductivistic conception of existence itself as a merely logico-linguistic matter. Rather it is to be interpreted in practical terms: to determine what the ontological commitments of a scientific theory are, it is necessary to determine the values of the quantified variables used in its canonical formalization.

Quine's approach is thus most properly conceived not as a reduction of ontology to the study of scientific language, but rather as a continuation of ontology in the traditional sense. When viewed in this light, however, it can be seen to be in need of vital supplementation. For the objects of scientific theories are discipline-specific. This means that the *relations* between objects belonging to different disciplinary domains fall out of bounds for Quinean ontology. Only something like a *philosophical* theory of how different scientific theories (or their objects) relate to each other can fulfill the task of providing an inventory of all the types of entities in reality. Quine himself would resist this latter conclusion. For him the best we can achieve in ontology lies in the quantified statements of particular theories, theories supported by the best evidence we can muster. We have no way to rise above the particular theories we have; no way to harmonize and unify their respective claims.

Internal vs. External Metaphysics

Quine is a realist philosopher. He believes in a world beyond language and beliefs, a world which the theories of natural science give us the power to illuminate. There is, however, another tendency in twentieth-century analytic philosophy, a tendency often associated with Quine but inspired much rather by Kant and promulgated by thinkers such as Carnap and Putnam. According to these thinkers ontology is a metalevel discipline which concerns itself not with the world itself but rather only with theories or languages or systems of beliefs. Ontology as a first-level science of reality – ontology as what these philosophers call "external metaphysics" – is impossible. The best we can achieve, they hold, is *internal metaphysics*, which means precisely the study of the ontological commitments of specific theories or systems of beliefs. Strawsonian descriptive metaphysics is one example of such internal metaphysics. Model-theoretic semantics, too, is often implicitly understood in internal-metaphysical terms – the idea being that we cannot understand what a given language or theory is really about, but we can build *models* with more or less nice properties. What we can never do is compare these models to some reality beyond. Ontology in the traditional philosophical sense thus comes to be replaced by the study of how a given language or science conceptualizes a given domain. It becomes a theory of the ontological content of certain representations. Traditional ontologists are seeking principles that are true of reality. The practitioners of internal metaphysics, in contrast, are seeking to elicit principles from subjects or theories. The elicited principles may or may not be true, but this, to the practitioner of internal metaphysics, is of no concern, since the significance of these principles

lies elsewhere – for instance in yielding a correct account of the taxonomical system used by speakers of a given language or by the scientists working in a given discipline.

In a development that has hardly been noted by philosophers, a conception of the job of the ontologist close to that of Carnap and Putnam has been advanced in recent years also in certain extraphilosophical disciplines, as linguists, psychologists, and anthropologists have sought to elicit the ontological commitments ("ontologies," in the plural) of different cultures and groups. Thus, they have sought to establish the ontology underlying common-sense or folk theories of various sorts by using the standard empirical methods of the cognitive sciences (see for example Keil 1979, Spelke 1990). Researchers in psychology and anthropology have sought to establish what individual human subjects, or entire human cultures, are committed to, ontologically, in their everyday cognition, in much the same way in which Quine-inspired philosophers of science had attempted to elicit the ontological commitments of the natural sciences.

It was still reasonable for Quine to identify the study of ontology – the search for answers to the question: what exists? – with the study of the ontological commitments of natural scientists. This is because it is a reasonable hypothesis that all natural sciences are in large degree consistent with each other. Moreover, the identification of ontology with ontological commitments continues to seem reasonable when one takes into account not only the natural sciences but also certain commonly shared commitments of common sense – for example that *tables* and *chairs* and *people* exist. For the common-sense taxonomies of objects such as these are compatible with those of scientific theory if only we are careful to take into account the different granularities at which each operates (Forguson 1989, Omnès 1999, Smith & Brogaard 2001).

Crucially, however, the running together of ontology and ontological commitments becomes strikingly less defensible when the ontological commitments of various specialist groups of *non*scientists are allowed into the mix. How, ontologically, are we to treat the commitments of astrologists, or clairvoyants, or believers in leprechauns? We shall return to this question below.

Ontology and Information Science

In a related development, also hardly noticed by philosophers, the term "ontology" has gained currency in recent years in the field of computer and information science (Welty & Smith 2001).

The big task for the new "ontology" derives from what we might call the Tower of Babel problem. Different groups of data- and knowledge-base system designers have their own idiosyncratic terms and concepts by means of which they build frameworks for information representation. Different databases may use identical labels but with different meanings; alternatively the same meaning may be expressed via different names. As ever more diverse groups are involved in sharing and translating ever more diverse varieties of information, the problems standing in the way of putting this information together within a single system increase geometrically. Methods must be found to resolve the terminological and conceptual incompatibilities which then inevitably arise.

Initially, such incompatibilities were resolved on a case-by-case basis. Gradually, however, it was recognized that the provision, once and for all, of a common reference ontology – a shared taxonomy of entities – might provide significant advantages over such case-by-case resolution, and the term "ontology" came to be used by information scientists to describe the construction of a canonical description of this sort. An ontology is in this context a dictionary of terms formulated in a canonical syntax and with commonly accepted definitions designed to yield a lexical or taxonomical framework for knowledge representation which can be shared by different information-systems communities. More ambitiously, an ontology is a formal theory within which not only definitions but also a supporting framework of axioms is included (perhaps the axioms themselves provide implicit definitions of the terms involved).

The methods used in the construction of ontologies thus conceived are derived on the one hand from earlier initiatives in database management systems. But they also include methods similar to those employed in philosophy (as described in Hayes 1985), including the

methods used by logicians when developing formal semantic theories.

Upper-level Ontologies

The potential advantages of ontology thus conceived for the purposes of information management are obvious. Each group of data analysts would need to perform the task of making its terms and concepts compatible with those of other such groups only once – by calibrating its results in the terms of the single canonical backbone language. If all databases were calibrated in terms of just one common ontology (a single consistent, stable, and highly expressive set of category labels), then the prospect would arise of leveraging the thousands of person-years of effort that have been invested in creating separate database resources in such a way as to create, in more or less automatic fashion, a single integrated knowledge base of a scale hitherto unimagined, thus fulfilling an ancient philosophical dream of a Great Encyclopedia comprehending all knowledge within a single system.

The obstacles standing in the way of the construction of a single shared ontology in the sense described are unfortunately prodigious. Consider the task of establishing a common ontology of world history. This would require a neutral and common framework for all descriptions of historical facts, which would require in turn that all legal and political systems, rights, beliefs, powers, and so forth, be comprehended within a single, perspicuous list of categories.

Added to this are the difficulties which arise at the point of adoption. To be widely accepted an ontology must be neutral as between different data communities, and there is, as experience has shown, a formidable trade-off between this constraint of neutrality and the requirement that an ontology be maximally wide-ranging and expressively powerful – that it should contain canonical definitions for the largest possible number of terms. One solution to this trade-off problem is the idea of a top-level ontology, which would confine itself to the specification of such highly general (domain-independent) categories as: time, space, inherence, instantiation, identity, measure, quantity, functional dependence, process, event, attribute, boundary, and so on. (See for example <http://suo.ieee.org>.) The top-level ontology would then be designed to serve as common neutral backbone, which would be supplemented by the work of ontologists working in more specialized domains on, for example, ontologies of geography, or medicine, or ecology, or law, or, still more specifically, ontologies of built environments (Bittner 2001), or of surgical deeds (Rossi Mori et al. 1997).

Uses of Ontology

The initial project of building one single ontology, even one single top-level ontology, which would be at the same time nontrivial and also readily adopted by a broad population of different information-systems communities, has however largely been abandoned. The reasons for this can be summarized as follows. The task of ontology-building proved much more difficult than had initially been anticipated (the difficulties being at least in part identical to those with which philosophical ontologists have grappled for some 2000 years). The information-systems world itself, on the other hand, is very often subject to the short time horizons of the commercial environment. This means that the requirements placed on information systems change at a rapid rate, so that already for this reason work on the construction of corresponding ontological translation modules has been unable to keep pace.

Yet work in ontology in the information-systems world continues to flourish, and the principal reason for this lies in the fact that its focus on classification (on analysis of object types) and on constraints on allowable taxonomies has proved useful in ways not foreseen by its initial progenitors. The attempt to develop terminological standards, which means the provision of explicit specifications of the meanings of the terms used in application domains such as medicine or air-traffic control, loses nothing of its urgency even when it is known in advance that the more ambitious goal of a common universal ontology is unlikely to be realized.

Consider the following example, due to Guarino. Financial statements may be prepared under either the US GAAP or the IASC standards (the latter being applied in Europe and many other countries). Cost items are often allocated to different revenue and expenditure categories under the two standards, depending on the tax laws and accounting rules of the countries involved. So far it has not been possible to develop an algorithm for the automatic conversion of income statements and balance sheets between the two systems, since so much depends on highly volatile case law and on the subjective interpretation of accountants. Not even this relatively simple problem has been satisfactorily resolved, though this is *prima facie* precisely the sort of topic where ontology could contribute something of great commercial impact.

If Ontek did not Exist, it would be Necessary to Invent It

Perhaps the most impressive attempt to develop an ontology – at least in terms of sheer size – is the CYC project (http://www.cyc.com), which grew out of an effort initiated by Doug Lenat in the early 1980s to formalize common-sense knowledge in the form of a massive database of axioms covering all things, from governments to mothers. The resultant ontology has been criticized for what seems to be its lack of principle in the ways in which new terms and theories come to be added to the edifice of the theory. CYC takes the form of a tangled hierarchy, with a topmost node labeled *Thing*, beneath which are a series of cross-cutting total partitions, including: *Represented Thing* vs. *Internal Machine Thing*, *Individual Object* vs. *Collection*, *Intangible* vs. *Tangible Object* vs. *Composite Tangible and Intangible Object*. Examples of Intangible Objects (*Intangible* means: *has no mass*) are sets and numbers. A person, in the CYC ontology, is a Composite Object made up of a Tangible body and an Intangible mind.

More important for our purposes here is the work of the firm Ontek – short for "ontological technology" – which since 1981 has been developing database programming and knowledge representation technologies necessary to create decision automation systems – "white collar robots" – for large-scale industrial enterprises in fields such as aerospace and defense. Realizing that the ontology required to build such systems would need to embrace in a principled way the entire gamut of entities encompassed by these businesses in a single, unified framework, Ontek approached this problem by systematically exploiting the resources of ontology in the traditional (adequatist) philosophical sense. A team of philosophers (including David W. Smith and Peter Simons) collaborated with software engineers in constructing the system PACIS (for Platform for the Automated Construction of Intelligent Systems), which is designed to implement a comprehensive theory of entities, ranging from the very concrete (aircraft, their structures, and the processes involved in designing and developing them), to the somewhat abstract (business processes and organizations, their structures, and the strategies involved in creating them), to the exceedingly abstract formal structures which bring all of these diverse components together.

Ontek has thus realized in large degree the project sketched by Hayes in his "Naïve Physics Manifesto," of building a massive formal theory of (in Hayes's case) common-sense physical reality (in Ontek's case this is extended to include not only airplane wings and factories but also associated planning and accounting procedures). As Hayes insisted, if long-term progress in artificial intelligence is to be achieved it is necessary to put away the toy worlds of classical AI research and to concentrate instead on the task of formalizing the ontological features of the world itself, as this is encountered by adults engaged in the serious business of living.

The Leipzig project in medical ontology (see <http://ifomis.de>), too, is based on a realist methodology close to that of Ontek, and something similar applies also to the work of Guarino and his colleagues in Italy. Most prominent information-systems ontologists in recent years, however, have abandoned the Ontologist's Credo and have embraced instead a view of ontology as an inwardly directed discipline (so that they have in a sense adopted an epistemologized reading of ontology analogous to that of Carnap and

Putnam). They have come to hold that ontology deals not with reality itself but rather with "alternative possible worlds," worlds which are indeed defined by the information systems themselves. This means not only that only those entities exist which are represented in the system (Gruber 1995), but also that the entities in question are allowed to possess only those properties which the system itself can recognize. It is as if Hamlet, whose hair (we shall suppose) is not mentioned in Shakespeare's play, would be not merely neither bald nor nonbald, but would somehow *have no properties at all* as far as hair is concerned. (Compare Ingarden 1973 on the "loci of indeterminacy" within the stratum of represented objects of a literary work.) What this means, however, is that the objects represented in the system (for example, people in a database) are not real objects – the objects of flesh and blood we find all around us. Rather, they are denatured surrogates, possessing only a finite number of properties (sex, date of birth, social security number, marital status, employment status, and the like), and being otherwise entirely indeterminate with regard to those sorts of properties with which the system is not concerned.

Information-systems ontologies in the sense of Gruber are, we see, not oriented around the world of objects at all. Rather, they are focused on our concepts or languages or mental models (or, on a less charitable interpretation, the two realms of objects and concepts are simply confused). It is in this light that we are to interpret passages such as the following:

> an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy. (Gruber, n.d.)

Conceptualizations

The newly fashionable usage of "ontology" as meaning just "conceptual model" is by now firmly entrenched in many information-systems circles. Gruber is to be given credit for having crystallized the new sense of the term by relating it to the technical definition of "conceptualization" introduced by Genesereth and Nilsson in their *Logical Foundation of Artificial Intelligence* (1987). In his 1993 article Gruber defines an ontology as "the specification of a conceptualization." Genesereth and Nilsson conceive conceptualizations as extensional entities (they are defined in terms of sets of relations), and their work has been criticized on the grounds that this extensional understanding makes conceptualizations too remote from natural language, where intensional contexts predominate (see Guarino, Introduction to 1998). For present purposes, however, we can ignore these issues, since we shall gain a sufficiently precise understanding of the nature of "ontology," as Gruber conceives it, if we rely simply on the account of conceptualizations which he himself gives in passages such as the following:

> A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly. (Gruber 1995)

The idea is as follows. As we engage with the world from day to day we participate in rituals and we tell stories. We use information systems, databases, specialized languages, and scientific instruments. We buy insurance, negotiate traffic, invest in bond derivatives, make supplications to the gods of our ancestors. Each of these ways of behaving involves, we can say, a certain conceptualization. What this means is that it involves a system of concepts in terms of which the corresponding universe of discourse is divided up into objects, processes, and relations in different sorts of ways. Thus in a religious ritual setting we might use concepts such as *salvation* and *purification*; in a scientific setting we might use concepts such as *virus* and *nitrous oxide*; in a story-telling setting we might use concepts such as *leprechaun* and *dragon*. Such conceptualizations are often tacit; that is, they are often not thematized in any systematic way. But tools can be developed

to specify and to clarify the concepts involved and to establish their logical structure, and in this way we are able to render explicit the underlying taxonomy. We get very close to the use of the term "ontology" in Gruber's sense if we define an ontology as the result of such clarification – as, precisely, the specification of a conceptualization in the intuitive sense described in the above.

Ontology now concerns itself not with the question of ontological realism, that is with the question whether its conceptualizations are *true of* some independently existing reality. Rather, it is a strictly pragmatic enterprise. It starts with conceptualizations, and goes from there to the description of corresponding domains of objects (also called "concepts" or "classes"), the latter being conceived as nothing more than nodes in or elements of data models devised with specific practical purposes in mind.

In very many cases the domains addressed by ontological engineers are themselves the products of administrative fiat. The neglect of truth to independent reality as a measure of the correctness of an ontology is then of little import. In such cases the ontologist is called upon merely to achieve a certain degree of adequacy to the specifications laid down by the client, striving as best he can to do justice to the fact that what the client says may fall short, for example, when measured in terms of logical coherence. Truth (or the lack of truth) can be a problem also in non-administrative domains. Bad conceptualizations abound (rooted in error, myth-making, hype, bad linguistics, or in the confusions of ill-informed "experts" who are the targets of knowledge-mining). Conceptualizations such as these may deal *only* with created (pseudo-) domains, and not with any transcendent reality beyond. They call for a quite different approach than is required in those areas – above all in the areas addressed by the natural sciences – where the striving for truth to independent reality is a paramount constraint. Yet this difference in question has hardly been noted by those working on information-systems ontology – and this gives us one clue as to why the project of a common reference ontology applicable in domains of many different types should thus far have failed.

Considered against this background the project of developing a top-level ontology begins to seem rather like the attempt to find some highest common denominator that would be shared in common by a plurality of true and false theories. Attempts to construct such an ontology must fail if they are made on the basis of a methodology which treats all application domains on an equal footing. Instead, we must find ways to do justice to the fact that the different conceptualizations which serve as inputs to ontology are likely to be not only of wildly differing quality but also mutually inconsistent.

What can Information Scientists Learn from Philosophical Ontologists?

As we have seen, some ontological engineers have recognized that they can improve their models by drawing on the results of the philosophical work in ontology carried out over the last 2000 years. This does not in every case mean that they are ready to abandon their pragmatic perspective. Rather, they see it as useful to employ a wider repertoire of ontological theories and frameworks and, like philosophers themselves, they are willing to be maximally opportunistic in their selection of resources for purposes of ontology-construction. Guarino and Welty (2000), for example, use standard philosophical analyses of notions such as identity, part, set-membership, and the like in order to expose inconsistencies in standard upper-level ontologies such as CYC, and they go on from there to derive metalevel constraints which all ontologies must satisfy if they are to avoid inconsistencies of the sorts exposed.

Given what was said above, it appears further that information ontologists may have sound *pragmatic* reasons to take the philosopher ontologist's traditional concern for truth more seriously still. For the very abandonment of the focus on mere conceptualizations and on conceptualization-generated object-surrogates may itself have positive pragmatic consequences.

This applies even in the world of administrative systems, for example in relation to the GAAP/IASC integration problem referred to above. For ontologists are here working in a

type of theoretical context where they must move back and forth between distinct conceptualizations, and where they can find the means to link the two together only by looking at their common objects of reference in the real, flesh-and-blood world of human agents and financial transactions.

Where ontology is directed in this fashion, not towards a variety of more or less coherent surrogate models, but rather towards the real world of flesh-and-blood objects in which we all live, then this itself reduces the likelihood of inconsistency and systematic error in the theories which result; and, conversely, it increases the likelihood of our being able to build a single workable system of ontology that will be at the same time nontrivial. On the other hand, however, the ontological project thus conceived will take much longer to complete and it will face considerable internal difficulties along the way. Traditional ontology is a difficult business. At the same time, however, it has the potential to reap considerable rewards – not least in terms of a greater stability and conceptual coherence of the software artifacts constructed on its basis.

To put the point another way: it is precisely because good conceptualizations are transparent to reality that they have a reasonable chance of being integrated together in robust fashion into a single unitary ontological system. If, however, we are to allow the real world to play a significant role in ensuring the unifiability of our separate ontologies, then this will imply that those who accept a conceptualization-based methodology as a stepping stone towards the construction of adequate ontologies must abandon the attitude of tolerance towards both good and bad conceptualizations. It is this very tolerance which is fated to undermine the project of ontology itself.

Of course to zero-in on good conceptualizations is no easy matter. There is no Geiger-counter-like device which can be used for automatically detecting truth. Rather, we have to rely at any give stage on our best endeavors – which means concentrating above all on the work of natural scientists – and proceed in careful, critical, and fallibilistic fashion from there, hoping to move gradually closer to the truth via an incremental process of theory construction, criticism, testing, and amendment. As suggested in Smith and Mark (2001), there may be reasons to look beyond natural science, above all where we are dealing with objects (such as societies, institutions, and concrete and abstract artifacts) existing at levels of granularity distinct from those which readily lend themselves to natural-scientific inquiry. Our best candidates for good conceptualizations will, however, remain those of the natural sciences – so that we are, in a sense, brought back to Quine, for whom the job of the ontologist coincides with the task of establishing the ontological commitments of scientists, and of scientists alone.

What Can Philosophers Learn from Information-systems Ontologists?

Developments in modal, temporal, and dynamic logics as also in linear, substructural, and paraconsistent logics have demonstrated the degree to which advances in computer science can yield benefits in logic – benefits not only of a strictly technical nature, but also sometimes of wider philosophical significance. Something similar can be true, I suggest, in relation to the developments in ontological engineering referred to above. These developments can first of all help to encourage existing tendencies in philosophical ontology (nowadays often grouped under the heading "analytic metaphysics") towards opening up new domains of investigation, for example the domain of social institutions (Mulligan 1987, Searle 1995, Smith 2002), of patterns (Johansson 1998), of artifacts (Dipert 1993, Simons & Dement 1996), of boundaries (Smith 2001), of dependence and instantiation (Mertz 1996), of holes (Casati & Varzi 1994), and parts (Simons 1987). Secondly, it can shed new light on the many existing contributions to ontology, from Aristotle to Goclenius and beyond (Burkhardt & Smith 1991), whose significance was for a long time neglected by philosophers in the shadow of Kant and other enemies of metaphysics. Thirdly, if philosophical ontology can properly be conceived as a kind of generalized chemistry, then

information systems can help to fill one important gap in ontology as it has been practiced thus far, which lies in the absence of any analog of chemical experimentation. For one can, as C. S. Peirce remarked (1933: 4.530), "make exact experiments upon uniform diagrams." The new tools of ontological engineering might help us to realize Peirce's vision of a time when operations upon diagrams will "take the place of the experiments upon real things that one performs in chemical and physical research."

Finally, the lessons drawn from information-systems ontology can support the efforts of those philosophers who have concerned themselves not only with the development of ontological *theories*, but also – in a field sometimes called "applied ontology" (Koepsell 1999, 2000) – with the *application* of such theories in domains such as law, or commerce, or medicine. The tools of philosophical ontology have been applied to solve practical problems, for example concerning the nature of intellectual property or concerning the classification of the human fetus at different stages of its development. Collaboration with information-systems ontologists can support such ventures in a variety of ways, first of all because the results achieved in specific application domains can provide stimulation for philosophers, but also – and not least importantly – because information-systems ontology is itself an enormous new field of practical application that is crying out to be explored by the methods of rigorous philosophy.

Acknowledgments

This chapter is based upon work supported by the National Science Foundation under Grant No. BCS-9975557 ("Ontology and Geographic Categories") and by the Alexander von Humboldt Foundation under the auspices of its Wolfgang Paul Program. Thanks go in addition to Thomas Bittner, Charles Dement, Andrew Frank, Angelika Franzke, Wolfgang Grassl, Nicola Guarino, Kathleen Hornsby, Ingvar Johansson, Kevin Mulligan, David W. Smith, William Rapaport, Chris Welty, and Graham White for helpful comments. They are not responsible for any errors which remain.

Bibliography

Bittner, Thomas. 2001. "The qualitative structure of built environments." *Fundamenta Informaticae* 46: 97–126. [Uses the theory of fiat boundaries to develop an ontology of urban environments.]

Brentano, Franz. 1981. *The Theory of Categories*. The Hague/Boston/London: Martinus Nijhoff. [Defends a classification of entities, and a new mereological view of substances and their accidents, based on Aristotle.]

Burkhardt, Hans and Smith, Barry, eds. 1991. *Handbook of Metaphysics and Ontology*, 2 vols. Munich/Philadelphia/Vienna: Philosophia. [Reference work on philosophical ontology and ontologists.]

Casati, Roberto and Varzi, Achille C. 1994. *Holes and Other Superficialities*. Cambridge, MA: MIT Press. [On the ontology and cognition of holes.]

Chisholm, Roderick M. 1996. *A Realistic Theory of Categories: An Essay on Ontology*. Cambridge: Cambridge University Press. [Defends a classification of entities based on Aristotle and Brentano.]

Dement, Charles W., Mairet, Charles E., Dewitt, Stephen E., and Slusser, Robert W. 2001. *Mereos: Structure Definition Management Automation (MEREOS Program Final Technical Report)*, available through the Defense Technical Information Center (http://www.dtic.mil/), Q3 2001. [Example of Ontek work in public domain.]

Dipert, Randall R. 1993. *Artefacts, Art Works and Agency*. Philadelphia: Temple University Press. [Defends a view of artifacts as products of human intentions.]

Forguson, Lynd. 1989. *Common Sense*. London/New York: Routledge. [Survey of recent work on common sense as a universal of human development and on the relevance of this work for the philosophy of common-sense realism.]

Genesereth, Michael R. and Nilsson, L. 1987. *Logical Foundation of Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann. [On logic in AI. Contains definition of "conceptualization" in extensionalist terms.]

Gruber, T. R. 1993. "A translation approach to portable ontology specifications." *Knowledge Acquisition* 5: 199–220. [Account of the language *ontolingua* as an attempt to solve the portability problem for ontologies.]

——. 1995. "Toward principles for the design of ontologies used for knowledge sharing." *International Journal of Human and Computer Studies*

43(5/6): 907–28. [An outline of motivations for the development of ontologies.]

——. [n.d.] "What is an Ontology?" <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.> [Summary statement of Gruber's definition of ontology as a specification of a conceptualization.]

Guarino, Nicola. 1995. "Formal ontology, conceptual analysis and knowledge representation." *International Journal of Human-Computer Studies* 43: 625–40. [Arguments for the systematic introduction of formal ontological principles in the current practice of knowledge engineering.]

——, ed. 1998. *Formal Ontology in Information Systems.* Amsterdam: IOS Press (Frontiers in Artificial Intelligence and Applications). [Influential collection.]

——. 1999. "The role of identity conditions in ontology design." In C. Freksa and D. M. Mark, eds., *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science.* Berlin/New York: Springer Verlag, pp. 221–34. [On constraints on ontologies.]

—— and Welty, C. 2000. "A formal ontology of properties." In R. Dieng and O. Corby, eds., *Knowledge Engineering and Knowledge Management: Methods, Models and Tools. 12th International Conference* (EKAW 2000). Berlin/New York: Springer Verlag, pp. 97–112. [A common problem of ontologies is that their taxonomic structure is poor and confusing as a result of the unrestrained use of subsumption to accomplish a variety of tasks. The paper provides a framework for solving this problem.]

Hayes, Patrick J. 1985. "The second naive physics manifesto." In J. R. Hobbs and R. C. Moore. eds., *Formal Theories of the Common-sense World.* Norwood, NJ: Ablex, pp. 1–36. [Against toy worlds methods in AI, in favor of massive axiomatization of common-sense physics along lines similar to much current work in ontology.]

Ingarden, Roman. 1964. *Time and Modes of Being*, tr. Helen R. Michejda. Springfield, IL: Charles Thomas. [Translated extracts from a masterly four-volume work in realist ontology entitled *The Problem of the Existence of the World* (the original published in Polish and in German).]

——. 1973. *The Literary Work of Art: An Investigation on the Borderlines of Ontology, Logic, and Theory of Literature*. Evanston, IL: Northwestern University Press. [Ontology applied to cultural objects.]

Johansson, Ingvar. 1989. *Ontological Investigations: An Inquiry into the Categories of Nature, Man and Society.* New York and London: Routledge. [Wide-ranging study of realist ontology.]

——. 1998. "Pattern as an ontological category." In Guarino ed. 1998: 86–94. [Ontology of patterns.]

Keil, Frank. 1979. *Semantic and Conceptual Development: An Ontological Perspective.* Cambridge, MA: Harvard University Press. [Study of cognitive development of category knowledge in children.]

Koepsell, David R., ed. 1999. *Proceedings of the Buffalo Symposium on Applied Ontology in the Social Sciences* (*The American Journal of Economics and Sociology*, 58[2]). [Includes studies of geographic ontology, the ontology of economic objects, the ontology of commercial brands, the ontology of real estate, and the ontology of television.]*****

——. 2000. *The Ontology of Cyberspace: Law, Philosophy, and the Future of Intellectual Property.* Chicago: Open Court. [A contribution to applied legal ontology with special reference to the opposition between patent and copyright.]

Mertz, D. W. 1996. *Moderate Realism and Its Logic.* New Haven, CT: Yale University Press. [Study of the logic and ontology of instantiation.]

Mulligan, Kevin, 1987. "Promisings and other social acts: their constituents and structure." In Kevin Mulligan, ed., *Speech Act and Sachverhalt. Reinach and the Foundations of Realist Phenomenology.* Dordrecht/Boston/Lancaster: D. Reidel, pp. 29–90. [On the ontology of speech acts as a foundation for a general ontology of social institutions.]

Omnès, Roland. 1999. *Understanding Quantum Mechanics.* Princeton: Princeton University Press. [Introduction to the consistent histories interpretation of quantum mechanics.]

Peirce, C. S. 1933. *Collected Papers.* Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1953. "On what there is," repr. in his *From a Logical Point of View*. New York: Harper & Row. [Defends a view of ontology as the study of the ontological commitments of natural science.]

Rossi Mori, A., Gangemi, A., Steve, G., Consorti, F., and Galeazzi, E. 1997. "An ontological analysis of surgical deeds." In C. Garbay et al., eds., *Proceedings of Artificial Intelligence in Europe* (AIME '97). Berlin: Springer Verlag.

Searle, John R. 1995. *The Construction of Social Reality*. New York: Free Press. [An ontology of social reality as the product of collective intentionality.]

Simons, Peter M. 1987. *Parts: An Essay in Ontology*. Oxford: Clarendon Press. [Logical and philosophical study of mereology.]

—— and Dement, Charles W. 1996. "Aspects of the mereology of artifacts." In Roberto Poli and Peter Simons, eds., *Formal Ontology*. Dordrecht: Kluwer, pp. 255–76.

Smith, Barry. 2001. "Fiat objects." *Topoi* 20(2). [On fiat and bona fide boundaries, with illustrations in geography and other domains.]

——. 2002. "John Searle: from speech acts to social reality." In B. Smith, ed., *John Searle*. Cambridge: Cambridge University Press.

—— and Brogaard, Berit. 2002. "Quantum mereotopology." *Annals of Mathematics and Artificial Intelligence* 7: 591–612. [A theory of the relations between partitions at different levels of granularity.]

—— and Mark, David M. 2001. "Geographical categories: an ontological investigation." *International Journal of Geographic Information Science* 15(7). [Study of naive subjects' conceptualizations of the geographical realm.]

Spelke, Elizabeth S. 1990. "Principles of object perception." *Cognitive Science* 14: 29–56. [On the role of objects (as the cohesive bounded foci of local action) in human cognitive development.]

Welty, C. and Smith, B., eds. 2001. *Formal Ontology and Information Systems*. New York: ACM Press. [Articles survey current work in information systems ontology.]

Chapter 12

Virtual Reality

Derek Stanovsky

Introduction

"Virtual reality" (or VR) is a strangely oxymoronic term. "Virtual," with its sense of "not actual" is jarringly juxtaposed with "reality" and its opposing sense of "actual." Undoubtedly the term has gained such currency at least partly because of this intriguing provocation. "Virtual reality" is currently used to describe an increasingly wide array of computer-generated or mediated environments, experiences and activities ranging from the near ubiquity of video games, to emerging technologies such as tele-immersion, to technologies still only dreamed of in science fiction and only encountered in the novels of William Gibson or Orson Scott Card, on the Holodeck of television's *Star Trek*, or at the movies in *The Matrix* of the Wachowski brothers, where existing VR technologies make possible a narrative about imagined VR technologies. The term "virtual reality" covers all of this vast, and still rapidly expanding, terrain.

"Metaphysics" too is an expansive term (see for example Chapter 11, Ontology, and Chapter 13, The Physics of Information). Setting itself the enormous task of investigating the fundamental nature of being, metaphysics inquires into what principles may underlie and structure all of reality. Some questions about virtual reality from the perspective of metaphysics might be: What sort of reality is virtual reality? Does the advent of virtual reality mark an extension, revision, expansion, or addition to reality? That is, is virtual reality real? Or is virtual reality more virtual than real and, thus, not a significant new metaphysical problem itself? How else might the links between "reality" and "virtuality" be understood and negotiated? Perhaps even more importantly, do the possible metaphysical challenges presented by virtual reality necessitate any changes in existing metaphysical views, or shed any light on other metaphysical problems?

This chapter approaches some of these questions, focusing on three main issues within the tremendously open field of inquiry laying at the intersection of metaphysics with virtual reality. First, the technology of virtual reality, along with some of the issues arising from this technology, will be situated and examined within the Western philosophical tradition of metaphysics stretching from ancient to modern and postmodern times. Next, the issues raised by virtual reality for personal identity and the subject will be explored and examined, beginning with Cartesian subjectivity and moving through poststructuralist theories of the subject and their various implications for virtual reality. Finally, these metaphysical considerations and speculations will be brought to bear on the current economic realities of globalization and the emerging information economy, which have become inextricably bound up with both

the metaphysics and politics of virtual reality as it exists today.

Since metaphysics itself is one of the broadest subjects, it seems odd to restrict the discussion of virtual reality only to one of its narrower senses. Therefore, virtual reality too will be construed as broadly as possible, and not confined to any one particular technological implementation, either existing or imagined. However, the insights concerning virtual reality gleaned in this manner should also find application in many of its narrower and more restricted domains as well. One final qualification: since metaphysics inquires into the fundamental structures of reality, and since it is unclear at this stage how virtual reality is to be located within reality, it might be more appropriate if the present inquiry into the metaphysics of virtual reality were described instead as an exercise in "virtual metaphysics." It may be that what virtual reality requires is not so much a place within the history of Western metaphysics as it does a metaphysics all of its own.

Virtual Reality

Virtual reality has been described in a variety of ways. In one of the earliest book-length treatments of virtual reality, Howard Rheingold writes: "One way to see VR is as a magical window onto other worlds . . . Another way to see VR is to recognize that in the closing decades of the twentieth century, reality is disappearing behind a screen" (Rheingold 1991: 19). This framing of virtual reality is a useful one for our purposes in that it helps to clarify and highlight one of the central issues at stake. Does virtual reality provide us with new ways to augment, enhance, and experience reality, or does it undermine and threaten that reality? Virtual reality is equally prone to portrayals as either the bearer of bright utopian possibilities or dark dystopian nightmares, and both of these views have some basis to recommend them. Before exploring these issues further, it will be helpful to describe and explain the origins of virtual reality, what virtual reality is currently, and what it may become in the future.

Virtual reality emerged from an unlikely hybrid of technologies developed for use by the military and aerospace industries, Hollywood, and the computer industry, and was created within contexts ranging from the cold war to science fiction's cyberpunk subculture. The earliest forms of virtual reality were developed as flight simulators used by the US military and NASA to train pilots. This technology led to the head-mounted displays and virtual cockpit environments used by today's fighter pilots to control actual aircraft. Another source of VR lies in the entertainment industry's search for ever more realistic movie experiences beginning with the early Cinerama, stereo sound, and 3D movies, and leading to further innovations in the production of realistic images and audio. Add to this a whole host of developments in computer technology. For instance, computer-aided design programs, such as AutoCAD, made it possible to use computers to render and manipulate three-dimensional representations of objects, and graphical computer interfaces pioneered by Xerox and popularized by Apple and Microsoft have all but replaced text-based computer interfaces and transformed the way people interact with computers. All of these trends and technologies conspired to create the technology that has come to be known as "virtual reality" (for more on the genesis and genealogy of VR see Rheingold 1991 and Chesher 1994).

There is not, or at least not yet, any fixed set of criteria clearly defining virtual reality. In his book *The Metaphysics of Virtual Reality*, Michael Heim identifies a series of "divergent concepts currently guiding VR research" each of which "have built camps that fervently disagree as to what constitutes virtual reality" (Heim 1993: 110). The cluster of features considered in this section concern computer-generated simulations which are interactive, which may be capable of being shared by multiple users, may provide fully realistic sensory immersion, and which may allow for forms of telepresence enabling users to communicate, act, and interact over great distances. Although not all of these elements exist in every version of virtual reality, taken together, these features have come to characterize virtual reality.

At one end of the spectrum, technologies allowing interactions with any representation or

simulation generated by means of a computer are capable of being described as virtual reality. Thus, a video game simulation of Kung-Fu fighting, or the icons representing "documents" on a simulated computer "desktop" might both be cases where computers create a virtual reality with which people then interact in a variety of ways. What makes these candidates for virtual reality is not simply the fact that they are representations of reality. Paintings, photographs, television, and film also represent reality. Computer representations are different because people are able to interact with them in ways that resemble their interactions with the genuine articles. In short, people can make the computer simulations do things. This is something that does not happen with other forms of representation. This form of virtual reality can already be provided by existing computer technologies and is becoming increasingly commonplace.

At the other end of the spectrum lie technologies aimed at fuller sensory immersion. Head-mounted displays, datagloves, and other equipment translate body, eye, and hand movements into computer input and provide visual, audio, and even tactile feedback to the user. This type of virtual reality aims at being able to produce and reproduce every aspect of our sensory world, with users interacting with virtual reality in many of the same ways they interact with reality, e.g. through looking, talking, listening, touching, moving, etc. (even tasting and smelling may find homes in virtual reality one day). Virtual reality in this vein aims at creating simulations that are not only perceptually real in how they look and sound, but also haptically and proprioceptively real in how they feel to users as well. As Randal Walser, a developer of virtual-reality systems, has written: "Print and radio tell; stage and screen show," while virtual reality "embodies" (quoted in Rheingold 1991: 192). At the imagined limit of such systems lie the virtual-reality machines of science fiction, with *Star Trek*'s Holodeck and the computer-generated world of *The Matrix* producing virtual realities that are perceptually and experientially indistinguishable from reality. No such technology exists today, but some elements of it are already possible.

In addition to the virtual reality of interactive simulations, whether confined to two-dimensional video screens, or realized through more ambitiously realistic and robustly immersive technologies, there are other elements that may also play a part in virtual reality. Perhaps the most important of these is provided by the capability of computers to be networked so that multiple users can share a virtual reality and experience and interact with its simulations simultaneously. The possibility for virtual reality to be a shared experience is one of the principal features by which virtual reality can be distinguished from fantasy. One of the tests of reality is that it be available intersubjectively. Thus, what is unreal about fantasy is not necessarily that the imagined experiences do not exist; it is that they do not exist for anyone else. Dreams are private experiences. On the contrary, the shared availability of virtual reality makes possible what William Gibson describes so vividly in his early cyberpunk novels of a computer-generated "consensual hallucination" (Gibson 1984: 51). The ability to share virtual reality sets the stage for a wide variety of human interactions to be transplanted into virtual reality, and opens opportunities for whole new avenues of human activity. Communication, art, politics, romance, and even sex and violence are all human activities that have found new homes in virtual reality. The possibility for the creation of entirely new forms of human interactions and practices that have no analog or precedence outside of virtual reality always remains open.

Another feature that may be encountered in virtual reality is that of "telepresence" or presence at distance, now frequently shortened simply to "presence." E-mail, video conferencing, distance education, and even telephones, all enable types of telepresence. In each of these cases, the technology allows users to communicate with distant people as if they were in the physical presence of each other. Such communication is so commonplace in so much of the world today, it hardly seems strange anymore that it is possible to communicate with people who are thousands of miles away. More sophisticated, realistic, and immersive technologies both exist, and can be imagined, that allow not only for written or spoken communication over great distances, but also for other types of interactions as well. For instance, the military use of remotely controlled aircraft and missiles, or the use of unmanned spacecraft

for exploration where humans might see, move, control, and use instruments to explore far-flung destinations in the solar system are both examples which allow human presence virtually. Other examples can be found in medicine, where surgeries are now performed via computer-controlled instruments, and surgeons interface with a video screen rather than a patient. These examples illustrate ways in which human presence, action, and interaction can be created virtually, and such examples are becoming more, rather than less, common.

Virtual reality not only creates new virtual spaces to inhabit and explore, but creates the possibility of virtual time as well. With the creation of computer-generated simulations came a bifurcation of time such that one now needs to distinguish between time in the simulated, virtual world and time in the rest of the world. Thus, only with the advent of the artificially created worlds of virtual reality does the concept of "real time" (RT) enter into general parlance. Communications and interactions in virtual reality (as opposed to IRL, "in real life") may be synchronous (as in video-conferencing and chatrooms) and coincide closely with real time, or asynchronous (as in e-mail exchanges) and diverge widely and unpredictably from the passage of time in other virtual interactions or with time outside the simulation. Time may even stop, or go backwards, within virtual reality. For instance, a simulation might be paused indefinitely, or reset to some previous state to allow users to experience a part of a simulation again. Time may also vary simply as a result of the technology used. This might happen when faster machines are networked with computers operating at lower MHz, or utilizing slower modems. In such cases, this can mean that some objects are rendered faster and changed and updated more frequently than others, giving an oddly disjointed sense of time, as objects in the same simulation move at distinctly different rates of time. These variations and complications in time emerge alongside and with virtual reality.

Not all of these elements exist in every version of virtual reality. However, taken together, they provide the background against which current virtual-reality systems are being invented and reinvented. These same elements also trace the horizon within which any metaphysics of virtual reality must take place.

Virtual Metaphysics

It is possible to recapitulate a large portion of the history of Western metaphysics from the vantage-point offered by virtual reality. The debates over rationalism, empiricism, realism, idealism, materialism, nominalism, phenomenology, possible worlds, supervenience, space, and time, to name just a few, can all find new purchase, as well as some new twists, in this brave new world of computer-generated virtual reality. This section traces some of the most influential Western metaphysical views concerning the distinction between appearance and reality and explores their possible relevance to virtual reality. This discussion by no means exhausts the metaphysical possibilities of virtual reality. In addition to the many strands of Western (henceforth this qualification will be omitted) metaphysics left untouched, there remain vast areas of metaphysical thought that could also be fruitfully explored, including long and rich traditions of African, Chinese, Indian, and Latin American metaphysics.

Distinguishing between appearance and reality is perhaps one of the most basic tasks of metaphysics, and one of the oldest, dating back at least to Thales and his pronouncement that despite the dizzying variety in how things appear, in reality "All is water." This desire to penetrate behind the appearances and arrive at the things themselves is one of the most persistent threads in metaphysics. Virtual reality presses at the very limits of the metaphysical imagination and further tangles and troubles long standing problems concerning how things seem versus how they really are. For instance, puzzles concerning mirrors and dreams and the ways in which they can confound our understanding of reality have a long history and haunt the writings of many metaphysicians. Virtual reality complicates these puzzles still further.

"But suppose the reflections on the mirror remaining and the mirror itself not seen, we would never doubt the solid reality of all that appears" (III. 6 [13]). This passage from Plotinus

comes wonderfully close to describing the current possibilities of virtual reality. Virtual reality may be very like the images in a mirror persisting even after the mirror disappears. In the case of mirrors, such a possibility remains only hypothetical. Plotinus assumes that in most cases the difference between reality and the reflection of reality presented by a mirror is easy to discern. After all, it is only Lewis Carroll's Alice who peers into a looking glass and takes what she sees to be a room "just the same as our drawing-room, only the things go the other way" (Carroll 1871: 141). Such a confusion seems amusingly childish and naive. So confident is Plotinus in this distinction between real objects and their unreal mirror images that he uses it as an analogy in support of his claim that reality lies with form rather than matter. However, what is more striking is that Plotinus allows that under certain circumstances (if the image in the mirror endured, and if the mirror itself was not visible) these reflections might fool us as well. Indeed, it is our inability to distinguish image from reality that lends interest to such spectacles as fun houses, with their halls of mirrors, and the illusions performed by magicians. In these cases, we do make the same mistake as Alice. It is this possibility of fundamentally conflating image, or representation, with reality that lends mirrors their metaphysical interest.

Virtual reality may present us with a new sort of mirror; one with the potential to surpass even the finest optical mirrors. If so, then virtual reality may fatally complicate the usual mechanisms used to distinguish image from reality, and representation from what is represented. For Plotinus, it is the limitations of the mirror image that reveals its status as a reflection of reality. It is only because images in a mirror are transient (fleeting, temporary, failing to persist over time or cohere with the rest of our perceptions) and because the mirror itself does not remain invisible (its boundaries glimpsed, or reflecting surface flawed or otherwise directly perceptible) that enables us to tell the difference between image and reality. One of the inherent limitations of any mirror is that it is necessarily confined to optical representations. Reaching out to touch an object in a mirror always reveals the deception. However, in immersive versions of virtual reality, the image need not be limited to sight. In virtual reality, the representation may pass scrutiny from any angle using any sense. As for transience, while the images in virtual reality may disappear at any moment, they also may be just as permanent and long-lived as any real object or event. Moreover, mirrors can only reflect the images of already existing things. Virtual reality has no such constraint. Objects in virtual reality may be copies of other things, but they also may be their own unique, individual, authentic objects existing nowhere else. This last point means that the grounds for needing to distinguish image from reality have changed. It is not simply that the representations of virtual reality are false (not genuine) like the reflections in a mirror. It is not even analogous to Plato's view of theater, which was to be banned from his Republic because of its distortions and misrepresentations of reality. Instead, virtual reality may summon up a whole new reality, existing without reference to an external reality, and requiring its own internal methods of distinguishing true from false, what is genuine or authentic from what is spurious or inauthentic.

Dreams too can provide occasions where perception and reality become interestingly entangled and may be one of the best, and most familiar, comparisons for virtual reality. Dreams possess many (although not all) of the elements of virtual reality. Dreams are immersive, matching in sensory clarity and distinctness even the most optimistic science fiction accounts of virtual reality. In his *Meditations*, Descartes famously entertains the possibility that there may be no certain method for distinguishing dreams from reality. He writes: "How often, asleep at night, am I convinced of just such familiar events – that I am here in my dressing gown, sitting by the fire – when in fact I am lying undressed in bed!" and finds such anecdotes sufficiently persuasive to conclude that "I see plainly that there are never any sure signs by means of which being awake can be distinguished from being asleep" (Descartes 1641: 77). Here, Descartes seems to suggest that dreams and reality can actually be confused, unlike Plotinus, who viewed the confusion of images in a mirror with reality as only a hypothetical possibility at best. Descartes, however, is unwilling to allow this

much uncertainty into his philosophical system and so appends the following curious solution to the dream problem in the last paragraph of his last Meditation. "But when I distinctly see where things come from and where and when they come to me, and when I can connect my perceptions of them with the whole of the rest of my life without a break, then I am quite certain that when I encounter these things I am not asleep but awake" (Descartes 1641: 122). Along with clarity and distinctness, Descartes adds coherence as a final criterion for certainty, in an effort to resolve the doubts raised by the dream problem. This is despite the fact that one of the chief strengths of the dream problem, as he put it forward, lay in the fact that dreams often could be fit coherently into waking life.

Virtual reality also can pass these tests of clarity, distinctness, and coherence. Beyond this, VR, unlike a dream, is able to satisfy the requirement of intersubjective availability that only "real" reality is generally assumed to possess. That is, whereas a dream can only be experienced by a single person, virtual reality is available to anyone. At this point, Descartes's dream problem takes on new life. Just as was true of the comparison with images in a mirror, the need to distinguish virtual reality from nonvirtual reality seems to dissolve. If virtual reality is not "real," it must be on some basis other than those considered so far. Distinguishing dream from reality, for Descartes, just like distinguishing image from reality for Plotinus, takes on importance precisely because, without some reliable means of discrimination, such confusions run the risk of infecting an otherwise easily recognized reality with instances of unreality. This would render reality a concept of dubious usefulness, for it could no longer clearly be distinguished from its opposite, from the unreal, from appearance, from image, or from dream. Descartes and Plotinus both identify permanence and coherence as criteria of the real and transience as the mark of the merely apparent. However, such solutions work even less well in the case of virtual reality. At this point the name "virtual reality" starts to become justified. Virtual reality takes on an existence with a distinctly different character from dreams, images, and other mere representations.

Other metaphysical systems plot more subtle and complex relationships between appearance and reality. Kantian metaphysics occupies a pivotal place in the history of metaphysics providing, as it does, a continuation of important strands of debate from antiquity, the culmination of several disputes within the modern period, and the origin of many contemporary discussions in the field. Can the Kantian system help provide a more sophisticated description of the status of virtual reality?

Kant's transcendental idealism revolves around the view that things in themselves are unknowable in principle and that human knowledge is only of appearances. Just like Descartes, Kant holds that we are epistemically acquainted with only our own perceptions. However, unlike Descartes, for Kant perceptual objects are nothing other than these patterns of representation encountered by the mind. Thus, Kant believes it is possible to overcome the epistemological problems introduced by the division between appearance and reality. This is because, for Kant, the mind plays an active, constitutive role in structuring reality. Chief among these contributions are the intuitions of space and time. Space and time are not themselves "things" that are directly perceptible, and yet, it is impossible for human beings to experience objects outside of space and time. What this means, according to Kant, is that "Both space and time . . . are to be found only *in us*" (1781: A 373). In this way, Kant hopes to overcome the epistemological divide between empiricism and rationalism by restricting knowledge to objects of experience, while at the same time granting an active role to the mind in structuring that experience.

Given a Kantian view, the objects encountered in virtual reality may not pose any significantly new metaphysical challenges. Since things in themselves are never the direct objects of human knowledge, the fact that experiences in virtual reality fail to correspond to objects outside the mind in any simple, straightforward way is not necessarily a problem. Every object of human knowledge, whether actual or virtual, is nothing other than just such an organized collection of perceptual representations. This means that virtual reality can be admitted to the world of empirical human experience on more or less

equal footing with the more usual forms of experience. Another way of stating this might be that, for Kant, all experience is essentially virtual. It is not epistemic contact with, or knowledge of, things as they exist apart from the mind that ever characterizes any human experience. What is known is only how those things appear to the mind. Given this, the fact that virtual reality exists for the mind (and can be made to exist for more than one mind) is sufficient to qualify those experiences as "real." One may still need to exercise some care in using and applying the empirical knowledge gained by way of virtual reality. Likewise, inferences based on that knowledge must be confined to their appropriate domain. However, this holds true for any piece of empirical knowledge no matter how it is acquired.

Kantian metaphysics may also help explain why human interactions with computers have conjured up these strange new frontiers of virtual space and virtual time. If it is true, as Kant conjectures, that the mind cannot experience things outside of space and time, then any new experiences will also have to be fit within these schemas. Although the mind does not possess innate ideas or any other particular content, it does provide a formal structure that makes possible any experience of the world. Presumably, this remains true of computer-generated worlds as well. Once computer-mediated experiences become a technical possibility, the mind also structures, organizes, and interprets these experiences within the necessary framework. Thus, virtual reality may be a predictable artifact of the mind's ordering of these computer-generated experiences. Virtual space and virtual time may be the necessary forms of apprehension of virtual reality, just as space and time are necessary to the apprehension of reality. In the case of virtual reality, the claim that space and time are "found only *in us*" seems much less contentious. Given these possibilities and connections, virtual reality may turn out to provide a laboratory for the exploration of Kantian metaphysics.

At this point one may wish to retreat to the relative safety of a more thoroughgoing materialism, where what is real is only the circuits and wires that actually produce virtual reality. However, the cost of such a move comes at the expense of the reality of all experience. It is not just Descartes and Kant who find a need to accord an increased status to ideas and perception. Even in Heidegger's existentialist metaphysics there is always not only the object, but also the encounter of the object; and these two moments remain distinct, and distinctly important. This experiential aspect of virtual reality is something that invites a more phenomenologically oriented approach. It may be tempting to see virtual reality as a vindication of Platonist metaphysics, where the world of ideas is brought to fruition and the less-than-perfect world of bodies and matter can be left behind. Others argue that rather than demonstrating the truth of Platonic idealism, or marking the completion of the Cartesian project of separating the mind from the body, virtual reality instead illustrates the inseparability of mind from body and the importance of embodiment for all forms of human experience and knowledge. After all, even in the noncorporeal world of virtual reality, virtual bodies had to be imported, re-created, and imposed in order to allow for human interaction with this new virtual world. This tends to point to the necessity of embodiment as a precondition for, rather than an impediment to, experience and knowledge (see Heidt 1999).

There are many other possible approaches to the metaphysics of VR. For instance, Jean Baudrillard's theories of simulation and hyperreality seem readymade for virtual reality, pointing to a metaphysics where contemporary social reality could be understood as having already fallen prey to the order of simulation made increasingly available by virtual reality. From Baudrillard's vantage-point, simulations, like those of VR, mark the end of our ability to distinguish between appearance and reality, reducing everything to a depthless hyperreality (see Baudrillard 1983). Another possibility would be to follow Jacques Derrida's critique of the metaphysics of presence onto the terrain of virtual reality where the absence of presence could be marked in new, high-tech ways. However, rather than pursuing additional examples, at this point it is better to inquire into a different, although closely related, set of metaphysical problems concerning the identity of the self.

Virtual Identity

In addition to raising questions about the nature and status of external reality, virtual reality also raises difficult questions concerning the nature of the subject, or self. Despite the differences in the metaphysical views discussed up to this point, there is one area of general agreement. Whether Platonist, Cartesian, or Kantian in orientation, in all of these systems there is a shared notion of a unified, and unifying, subject whose existence provides a ground for knowledge, action, and personal identity. Such a conception of the subject has been complicated in recent years. In particular, poststructuralist accounts of a divided and contingent subject have raised questions about the adequacy of previous views. Virtual reality also complicates assumptions concerning a unified subject. The example discussed above of images in a mirror can be used again to approach these questions surrounding the subject, this time through the work of Jacques Lacan.

Lacan's influential formulation of the "mirror stage" pushes the notion of the knowing subject to its limits. Inverting traditional Cartesian epistemology, the subject, instead of being the first and most surely known thing, becomes the first misrecognized and misknown thing. This is an even more radical mistake than that made by Alice in her trip through the looking glass. At least when Alice looked in the mirror and saw a girl very much like herself, she still took it to be a different little girl and not herself. For Descartes, this would amount to a mistake in the one thing he thought he could be certain of, the *cogito*. Given Lacan's view, "I think, therefore I am" becomes an occasion for error when pronounced while looking into a mirror. In this case, the I of thinking can differ from the I of existing (the I of consciousness thinks, therefore the I in the mirror exists). Lacan reworks the slogan to read, "I think where I am not, therefore I am where I do not think" (Lacan 1977: 166). Such a formulation could never serve as Descartes's foundation for knowledge once this division is introduced within the subject.

This divide within the subject is precisely what is highlighted in Lacan's discussion of the mirror stage. Lacan writes: "We have only to understand the mirror stage *as an identification*, in the full sense that analysis gives to the term: namely, the transformation that takes place in the subject when he assumes an image" (Lacan 1977: 2). The subject is thus produced by an identification with an image, an image that is not the subject and yet which is mistaken to be identical with it. If identity is based on identifications, and identification is always an identification with something one is not, then one's identity will always be something that is at odds with itself. Elsewhere, Lacan explicitly relies on an example of a trick done with mirrors to illustrate the situation of the human subject. Here, the illusion of a vase filled with flowers is produced. For Lacan, it is the illusion of the self that is produced. (See figure 12.1.)

In the figure, the subject occupies the position of the viewer (symbolized by a barred S to re-emphasize this division which founds the subject), and the ego is represented by the virtual image of the inverted vase seen in the mirror. Lacan is proposing that a mistake worse than that made by Alice with the looking glass is not merely commonplace, but constitutive of human subjectivity. The self, emerging over time as the result of a series of identifications with others, is, like the image of the vase in the mirror, not actual but virtual.

Virtual reality compounds this dilemma. If in reality the subject is already not where it thinks itself to be, in virtual reality the situation becomes even worse. Virtual reality provides an open field for various and even multiple identities and identifications. In virtual environments, people are not confined to any one stable unifying subject position, but can adopt multiple identities (either serially or simultaneously). From the graphical avatars adopted to represent users in virtual environments, to the handles used in chatrooms, to something as simple as multiple e-mail accounts, all of these can be used to produce and maintain virtual identities. Identity in virtual reality becomes even more malleable than in real life, and can be as genuine and constitutive of the self as the latter. Sexual and racial identities can be altered, edited, fabricated, or set aside entirely. Identities can be ongoing, or adopted only temporarily. Thus, virtual reality opens the possibility not only of recreating space and time,

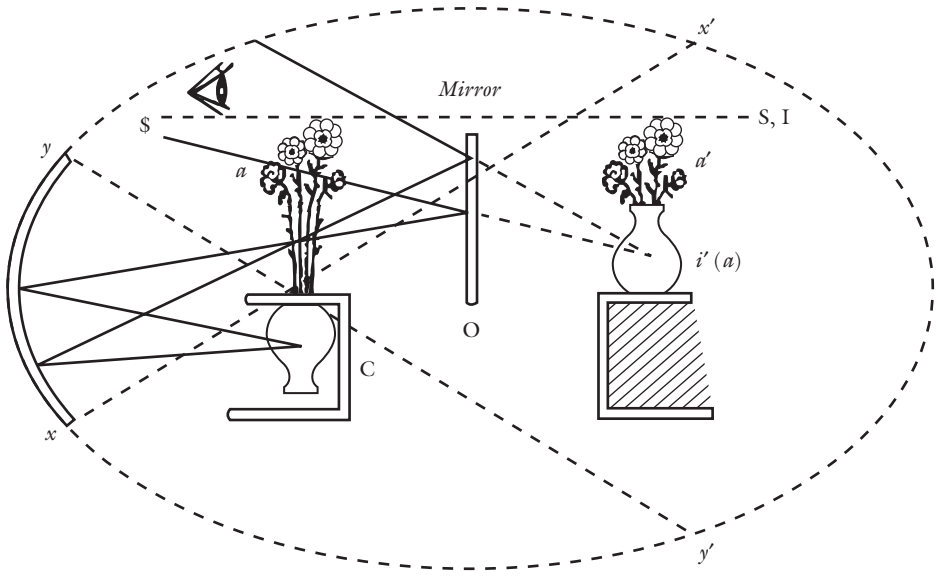

Figure 12.1: The illusion of a vase/the illusion of the self (Lacan 1978)
Source: "Diagram on p. 145," from *The Four Fundamental Concepts of Psycho-analysis* by Jacques Lacan, tr. Alan Sheridan. Copyright © 1973 by Editions du Seuil. English tr. copyright © 1977 by Alan Sheridan. Used by permission of W. W. Norton & Co. Inc.

but the self as well. The subject is produced anew as it comes to occupy this new space. In her influential book *Life on the Screen*, Sherry Turkle argues that online identities make "the Gallic abstractions" of French theorists like Lacan "more concrete," writing: "In my computer-mediated worlds, the self is multiple, fluid, and constituted in interaction with machine connections; it is made and transformed by language" (Turkle 1995: 15). For Turkle, the divisions and fragmentations that mark every identity take on new prominence and find new uses in the virtual reality of online society.

Economic Reality

The metaphysics of virtual reality may strike some as the most esoteric of topics, far removed from everyday life and practical human concerns. However, metaphysical views often have a surprising reach and can make their influence felt in unsuspected ways. In the case of virtual reality, these metaphysical attachments are currently in the process of producing and reshaping vast areas of our social reality. If virtual reality has yet to supplant more traditional modes of human interaction with the physical world, with each other, and even with oneself, there is one arena in which virtual reality has already made startling and astonishingly swift inroads, and that is in the realm of economics. From ATM machines and electronic transfers to the dot-com boom and bust, global capital has not been shy about leaping into the virtual world of e-commerce. Why has global capital been able to find a home in this new virtual economic space with such ease and rapidity? What does this colonization of virtual reality portend for other noncommodity possibilities of virtual reality?

Globalization is a process that has certainly been facilitated by the information economy of the digital age. Mark Poster has described this situation as "Capitalism's linguistic turn" as the industrial economy segued into the information economy (Poster 2001: 39). Capital has been instrumental in producing and disseminating the technologies that have made this process possible. The coining of the phrase "virtual reality"

is most often attributed to Jaron Lanier, a developer and entrepreneur of virtual-reality systems, to use as part of a marketing strategy for his software company. The potential of e-mail as an advertising medium was pioneered early on when, in 1994, a pair of enterprising green-card attorneys became the first to use e-mail as a form of direct marketing. Computer sales, driven by the expansion of the internet, fueled the expansion of the high-tech economy to such an extent that the internet service provider America Online could afford to buy media giant Time Warner. Virtual reality has created new commodities, which have quickly become new economic realities. Capital has also tended to transplant and reproduce already existing social and economic inequalities into this new virtual world. For instance, there has been much discussion of the "digital divide" between those with access to global information networks and those without. This divide falls along the well-worn demarcations of race and gender, but even more starkly, along class lines. The divide between rich and poor, both within and between nations, has been mapped onto the very foundations of the information age. These capitalist origins of virtual reality should not be forgotten.

Capital organizes economic and social life around the production and consumption of commodities. Marx writes that the commodity form raises a whole host of "metaphysical subtleties and theological niceties" (Marx 1867: 163). Relationships between commodities become "dazzling" in their variety and movements, while the social relationships between producers and consumers become obscured behind the appearances of wages and prices (Marx 1867: 139). For Marx, the value of a commodity only emerges virtually. The value of one commodity finds expression only in the body of another commodity through the relationship of exchange. Thus, the value of a watch might be expressed in its exchange for a cellphone. This system of exchange finds its culmination in money, a commodity whose function is to provide a mirror for the value of every other commodity. The particular commodity serving as money changes over time, from gold and silver to paper and plastic, as money asymptotically approaches the perfect mirror described by Plotinus, where only the image remains and the mirror disappears. The current electronic transfer of funds around the globe comes close to realizing this goal (for a further discussion of "digital gold" in the information age, see Floridi 1999: 113ff). It may be that this spectral nature of money means that capital is uniquely adapted for virtual reality. Money is already the virtual expression of value.

For capital, the additional "metaphysical subtleties" tacked on by virtual reality may scarcely matter. The already virtual existence of money has facilitated the migration of capital into virtual reality with nothing lost in the transition. The online virtual reality of the internet was once home to a variety of small, but close-knit, virtual communities. This has changed. Now the character and function of the internet more closely resembles a virtual shopping mall as advertisements appear everywhere and the identity of consumer overtakes every other online identity. We may currently be living through a process of virtual primitive accumulation, or a kind of electronic enclosure movement, as the free association and utopian possibilities offered by online virtual reality are driven out by the commodification imposed by global capital. Capital, long a kind of universal solvent for social relations, is currently transforming the virtual social relations of online life at a breathtaking pace. However, this process does not occur without active resistance (see Chesher 1994, and Dyer-Witheford 1999). It is here that the urgency of these otherwise abstract metaphysical speculations can be felt. The metaphysics of virtual reality provides the horizon on which a host of new ethical and political questions will take shape and within which they must be answered.

References

Baudrillard, J. 1983. *Simulations*, tr. P. Foss, P. Patton, and P. Beitchman. New York: Semiotext(e).

Carroll, L. 2000 [1871]. *The Annotated Alice: Alice's Adventures in Wonderland and Through the Looking-glass.* New York: W. W. Norton. Original works published 1865 and 1871.

Chesher, C. 1994. "Colonizing virtual reality: construction of the discourse of virtual reality,

1984–1992." *Cultronix* 1(1) (Fall 1994). <http://eserver.org/cultronix/chesher>.

Descartes, R. 1988 [1641]. *Descartes: Selected Philosophical Writings*, tr. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press. *Meditations on First Philosophy* originally published 1641.

Dyer-Witheford, N. 1999. *Cyber-Marx: Cycles and Circuits of Struggle in High-Technology Capitalism*. Urbana: University of Illinois Press. [This book provides a thorough and detailed account of the vicissitudes of class struggle within the global capitalist information economy from an autonomist Marxist perspective.]

Floridi, L. 1999. *Philosophy and Computing: An Introduction*. London: Routledge. [This textbook provides a clear and accessible overview of philosophical problems relating to computers and information theory ranging from the internet to artificial intelligence.]

Gibson, W. 1984. *Neuromancer*. New York: Ace Books.

Heidt, S. 1999. "Floating, flying, falling: a philosophical investigation of virtual reality technology." *Inquiry: Critical Thinking Across the Disciplines* 18(4): 77–98.

Heim, M. 1993. *The Metaphysics of Virtual Reality*. New York: Oxford University Press. [Michael Heim's book gives an accessible introduction to some of the philosophical issues arising from virtual reality and explores the changes this technology may have introduced into reality.]

Kant, I. 1996 [1781]. *Critique of Pure Reason: Unified Edition*, tr. W. S. Pluhar. Indianapolis: Hackett. Original works published 1781 and 1787.

Lacan, J. 1977. *Écrits: A Selection*, tr. A. Sheridan. New York: W. W. Norton.

——. 1978. *The Four Fundamental Concepts of Psycho-analysis*, tr. A. Sheridan. New York: W. W. Norton.

Marx, K. 1977 [1867]. *Capital, Volume One*, tr. B. Fowkes. New York: Vintage Books. Original work published 1867.

Plotinus. 1992 *The Enneads*, tr. S. MacKenna. Burdett, NY: Paul Brunton Philosophic Foundation.

Poster, M. 2001. *What's the Matter with the Internet?* Minneapolis: University of Minnesota Press. [Mark Poster's book provides a sophisticated inquiry into the culture and politics of the internet and covers topics on critical theory, postmodernism, globalization, and democracy.]

Rheingold, H. 1991. *Virtual Reality*. New York: Summit Books. [This book is a lively, and accessible, journalistic account of the people and history behind the early development of the technologies of virtual reality.]

Turkle, S. 1995. *Life on the Screen: Identity in the Age of the Internet*. New York: Touchstone. [Sherry Turkle's book is one of the new classics of internet studies. Drawing on psychoanalysis and French theory to explore online identities, Turkle examines the fragmented nature of the self in postmodern culture.]

Chapter 13

The Physics of Information

Eric Steinhart

1 Introduction

This chapter has two goals. The first is to analyze physical concepts like space, time, and causality in informational and computational terms (the informational/computational nature of physics). The other is to explain some key informational and computational concepts in physical terms (the physical nature of information/computation). These two goals are philosophically interesting for at least six reasons. (1) Philosophers have always been interested in the logical structure of physical reality, even metaphorically. The images of the physical universe as an arrangement of geometrical figures or a clock have been superseded by the image of the universe as a computer. Metaphors apart, we shall see that, strictly speaking, the universe is a computer if and only if its physics is *recursive*. (2) Claims about the computational powers of physical systems appear in many philosophical arguments. Some arguments in the philosophy of mind, for example, depend on the computational powers of physical systems like the human body and brain. (3) The role of the transfinite in the philosophical conception of computation requires clarification. If an idealized Turing machine has infinitely many squares on its tape, then it ought to be able to have infinitely many 1s or to make infinitely many moves. The calculus has long provided physical theory with an apparently consistent notion of limits and the theory of physical computations should be able to take advantage of that idea. Failure to consider infinities adequately has led many thinkers to regard finite Turing computability as some sort of necessary upper bound to physical computation. This would be a mistake. It is possible that there are physical hypercomputers far more powerful than classical Turing machines (even if none exist in our universe). (4) Philosophical efforts to analyze the mind–brain relation in terms of programs and computers (e.g. functionalism) seem to have introduced a kind of dualism between software and hardware. The ontology of software is unnecessarily vague. "Virtual" software objects are often described as if they were nonphysical objects that nevertheless participate in spatio-temporal-causal processes. Hence the "virtual" can become a strange category, neither concrete nor abstract. (5) Sometimes, philosophers make false claims about computers. One hears about continuous "analog" computers even though all the quantities involved are known to be discrete. Or one is told that computers manipulate information as if it were some kind of immaterial stuff (like pneuma or ether). Finally, (6) theoretical efforts to understand physical reality in computational terms are often confused with attempts to devise simulations of physical systems. However, physicists who wonder whether the universe is a

computer are not concerned with virtual reality models or simulations, they are concerned with the spatio-temporal-causal structure of the physical universe itself.

2 Programs and Theories

A *program* can be described as a recursive definition of some property *P*, consisting of at least two parts: a *basis* clause, which states some initial fact about *P*, and a *recursion* clause, which defines new facts about *P* in terms of old facts. Properties that have recursive definitions are simply referred to as "recursive." If *P* is recursive, then the set of all objects that have *P* is also described as recursive, and each object in that set is said to be a recursive object. Many physical things are recursive. For example, the property *is-a-chain* can be defined recursively thus: (1) a link O is a chain; (2) if X is a chain and O is a link, then X attached to O is a chain. The definition generates a series of chains: O, OO, OOO, etc., where each chain is a linear or one-dimensional series of neighboring points (the links). This recursive definition can be extended to generate discrete structures with more dimensions *D*. A *grid* is a 2D arrangement of neighboring points, i.e. a set of points and a recursive neighbor relation that determines a distance relation. All geometric facts about the grid are recursive. Here is an informal recursive definition of the property *is-a-grid*: (1) a set of four points occupying the corners of a square is a grid $G(0)$; (2) if $G(n)$ is a grid, then the set of points made by dividing each square in $G(n)$ into four equal squares is a grid $G(n + 1)$. Figure 13.1 shows the series of grids $G(0)$, $G(1)$, $G(2)$. The definition can be further extended to generate a 3D lattice whose points are at the corners of cubical cells. We can then add a fourth dimension. If this is time, the result is a 4D space-time structure in which all the spatial and temporal facts (distances and durations) are recursive.

A recursive definition associates each finite whole number with some set of facts. The basis clause associates 0 with some facts; the first application of the recursion clause associates 1 with some facts; the *n*th application associates n with some facts, and so forth. The recursive definition of is-a-chain, for example, associates 0 with O, 1 with OO, 2 with OOO, and so on. If *R* is some recursive definition, let $R(n)$ be the *n*th set of facts generated by *R*. So, $R(n)$ is the *n*th chain, or the *n*th grid. A set of facts is a *state of affairs*. A state of affairs *is finitely recursive* if and only if there is some recursive definition *R* and some finite whole number *n* such that the state of affairs is the set of facts $R(n)$.

To extend recursive definitions to the infinite one needs to define a state of affairs at infinity. The result is a definition by *transfinite recursion*. If some finite recursive definition *R* associates each finite number *n* with $R(n)$, then one way to extend *R* transfinitely is to add a *limit clause* that associates infinity (∞) with the limit of $R(n)$ as *n* goes to infinity. $R(\infty)$ is then the limit of $R(n)$ as *n* increases without bound. For example: recall Zeno's paradox of the racecourse, in which Achilles starts at 0 and runs to 1 by always going halfway first. Achilles traverses the points 1/2,

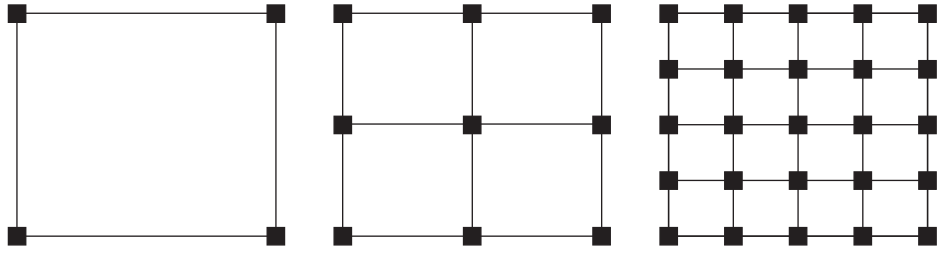

Figure 13.1: A recursive series of discrete 2D spaces

3/4, 7/8, and so on. The property *is-a-Zeno-point* is defined by finite recursion like this: (1) $R(0)$ is the Zeno point 1/2; (2) if $R(n)$ is the Zeno point P/Q, then $R(n + 1)$ is the Zeno point $P/Q + P/2Q$. This definition generates the series: 1/2, 3/4, 7/8, and so on. The theory of infinite series from the calculus shows that the limit of the series of Zeno points is 1. So we add the limit clause (3) $R(\infty)$ is the limit of $R(n)$ as n goes to ∞. The result is a transfinite recursive definition. If recursive definitions that take limits are available, then much of the calculus is also available. However, it is not necessary to restrict limit clauses to limits as defined in calculus. We can also let $R(\infty)$ be the state of affairs that contains all the facts in every $R(n)$ for which n is finite. $R(\infty)$ can be the union of all the $R(n)$ for n finite. We can thus define recursive states of affairs transfinitely. For instance, if the grid $G(\infty)$ is the union of all the $G(n)$ for n finite, then $G(\infty)$ is an infinitely subdivided space. It is a grid such that between any two points there is another. Generally, a state of affairs is *transfinitely recursive* if and only if there is some transfinite recursive definition R such that the state of affairs is $R(\infty)$.

A *theory* T can be described as a collection of facts that entails further facts. Every recursive definition is a theory. T is *ultimate* for some physical system S whenever T entails all and only the physical facts about S. A physical system S is recursive if and only if there is some recursive theory that is ultimate for S, i.e. there is some program that generates all and only the facts about S. If a physical system is *finitely recursive* this means that there is some recursive definition R and some finite number n such that $R(n)$ is ultimate for S; if it is *transfinitely recursive* this means that there is some recursive definition R and some transfinite number N such that $R(N)$ is ultimate for S; otherwise it is *nonrecursive*. It is known that there are entities whose definitions are nonrecursive. Nonrecursive systems typically involve the real numbers or logical undecidabilities. The three kinds: finitely recursive, transfinitely recursive, and nonrecursive are of course logically exhaustive. Since our universe is a physical system, it necessarily falls under one of those three kinds. Where our universe lies is an open question.

3 Finite Digital Physics

The set of logically possible programs is infinite and hence much larger than the set of programs that can actually be written by human beings. All programs have possible physical models. Since some of the programs we write are realized by artificial computers (which are physical parts of our universe), some programs are at least approximately true of some parts of our universe. Consider now that some states of affairs in our universe are finitely recursive, and that our universe is a large state of affairs, but it may be finite (Finkelstein 1995; Steinhart 1998). *Finite digital physics* argues that there is some recursive definition R and some finite number N, such that $R(N)$ entails all and only the physical facts about our universe. The recursive definition $R(N)$ is a program P that runs for N steps. Each step defines some change (some state-transition) of our universe. So finite digital physics suggests that there is some finite P that is exactly instantiated by our whole universe. P is the ultimate theory for our universe. We certainly cannot run P on any part of our universe; we may not be able to write P; what finite digital physics argues for is that P exists.

A physically possible universe U is *digital* if and only if it is finitely recursive. Since all finitely recursive quantities are digital, if U is finitely recursive then all its physical quantities are digital, i.e. discrete (measured by an integer variable) and with only finitely many values (finite upper and lower bounds). Discrete quantities are contrasted with dense quantities (measured by rational numbers) and continuous quantities (measured by real numbers). Since almost all variables in the classical (Newton–Maxwell) physical theory of our universe are real number variables that refer to continuous physical quantities, and since almost all the equations in classical physics are differential equations that refer to continuous transformations of those quantities, it might be argued that our universe is far too mathematically complex to be finitely recursive. However, differential equations may be over-idealizations. The Lotka–Volterra differential equations, for example, describe the interactions of predator–prey populations as if they were continuous, yet animals come in discrete units. Moreover,

classical physics has been replaced by quantum physics, where quantities (charge, mass, length, time) are quantized into discrete units (*quanta*). Thus, analytic work on the foundations of quantum physics motivates John Wheeler's theory (Zurek 1990) that physical reality emerges from binary alternatives. Wheeler himself has referred to this position by means of the slogan "Its from Bits." Zeilinger (1999) argues that the quantization of information (as bits) entails the quantization of all measurable physical quantities, and that this is a fundamental feature of physical systems. Whether all the fundamental physical quantities of our universe are digital is an open scientific question. Since it is a question about the *form* of the ultimate scientific theory, it is not clear whether there are experiments that can empirically decide this question.

If U is digital, then space-time has finitely many dimensions and it is finitely divided into atomic (0-dimensional) point-instants (cells). A digital space-time is a network of finitely many cells, in which each cell has links to finitely many spatial and temporal neighbors. Motion is not continuous, time proceeds in clock ticks, and space proceeds in steps. There is a maximal rate of change (a speed of light), namely one step per tick. Physical quantities in U are associated with geometrical complexes of cells (e.g. the 0-dimensional cells in U, the 1D links between cells, the 2D areas bounded by links, the 3D volumes bounded by areas, etc.). Each cell in a digital universe is associated with at most finitely many quantities, e.g. some digital mass. Links between cells are associated with digital spins or other forces; areas bounded by links are associated with digital charges.

Space and time in our universe could be finite and discrete. Relativity theory permits space to be finitely extended and, according to quantum mechanics, space is finitely divided with a minimal length (the Planck length, about 10^{-35} meters) and time has a minimal duration (the Planck time, about 10^{-43} seconds). Furthermore, theories of loop-quantum gravity predict that space is a discrete "spin-network."

Discrete space-times have mathematical features that may make them unsuitable for use as actual physical structures. First, discrete space-times have different geometries than continuous space-times. All distances in discrete space-times are integers, but a square whose sides have unit distance has a diagonal, and since the Pythagorean theorem shows that such diagonals are irrational (non-integer) numbers, the theorem cannot be true for discrete space-times. If we try to avoid this problem by treating the lengths of sides and the lengths of diagonals as fundamental lengths (of which we allow integer multiples), then we have two incommensurable distances. Space is no longer uniform. Second: discrete space-times have very limited internal symmetries for physical rotations and reflections. Square lattices allow only 90-degree rotations; triangular or hexagonal lattices allow only 60-degree rotations. Nevertheless, actual physics seems to demand rotations of any degree. Finally: it is difficult to translate continuous differential equations into discrete difference equations. It is not known whether these mathematical features are genuine obstacles, or whether they are merely inconveniences for scientists used to thinking of space-time as continuous. The study of discrete space-times is an active research area and it is likely that lattice quantum field theories will solve the problems associated with digital space-times. Whether our universe has digital space-time remains an open scientific question.

If U is digital, then the causal regularities in U are finitely recursive. The laws of nature are finitely recursive transformations of digital physical quantities and all the basic quantities stand to one another in finitely recursive arithmetical relations. Since all quantities are digital, an arithmetic transformation is possible (is consistent with the fact that U is digital) if and only if that operation takes some finitely bounded integers as inputs and produces some finitely bounded integers as outputs. More precisely: all the possible laws of nature for U must be recursive functions on the integers, and they must produce outputs within the finite upper and lower bounds on quantities in U. For example, suppose that U consists of a space that is a 2D grid, like a chessboard. Each cell on the board is associated with some quantity of matter (some mass either 0 or 1). The assignment of masses to cells is a discrete (binary) mass field. As time goes by (as the clock ticks), the mass field changes according to some causal operator. A causal operator for the mass

field of U is digital if and only if it defines the mass field at the next moment in terms of the mass field at some previous moments. The causal operator on a binary mass field is a Boolean function of bits that takes 0s and 1s as inputs and produces 0s and 1s as outputs. A *dynamical system* is a physical system whose states (synchronic distributions of quantities) and transitions (diachronic transformations of quantities) are recursively defined. It repeatedly applies a causal operator to its initial state to produce its next states. Such dynamical repetition or iteration is recursive change. If U is a digital universe and its causality is recursive and discrete, then U is a discrete dynamical system. Discrete dynamical systems are an area of active physical research. Our universe could be a discrete dynamical system, in which case, all differential equations that relate continuous rates of change are ultimately based on finite difference equations involving digital quantities in digital space-time. Whether our universe has digital causality is an open question.

There are many classes of digital universes. *Cellular automata* (CAs) are the most familiar digital universes. Conway's *Game of Life* is a popular cellular automaton (see Chapter 15, Artificial Life). Space, time, causality, and all physical quantities in CAs are finite and discrete. CAs are computational *field theories*: all quantities and transformations are associated with space-time cells. Causality in CAs is additionally constrained to be local: the quantities associated with each cell are recursively defined in terms of the quantities associated with the spatio-temporal neighbors of that cell. CA theory has seen great development (Toffoli & Margolus 1987), and CAs have seen extensive physical application (Chopard & Droz 1998). There are many generalizations of CAs: *lattice gasses* (Wolf-Gladrow 2000), *lattice quantum CAs* and *lattice quantum field theories* are currently active research areas in physics. Fredkin (1991) argues that our universe is a finitely complex CA.

4 Transfinite Digital Physics

Our universe may be too complex to be only *finitely recursive*. *Hyperdigital physics* argues that our universe is *transfinitely recursive*. If a universe U is *transfinitely recursive* this means that there is some recursive definition R and some transfinite N such that $R(N)$ is ultimate for U. The class of hyperdigital universes is very large. Hyperdigital physics permits physical infinities so long as they do not introduce logical inconsistencies. While finite recursion subdivides space-time finitely many times, transfinite recursion subdivides space-time infinitely. An infinitely subdivided space-time seems to be consistent. If U is finitely recursive, then each quantity in U is measured by only finitely many digits; but if U is transfinitely recursive, it can contain quantities measured by infinitely long series of digits. It is easy to define transformations on infinitely long series of digits by defining them in terms of endless repetitions of operations on single digits. A quantity measured by an infinitely long series of digits is infinitely precise. Infinitely precise physical arithmetic seems to be consistent, that is, it seems that physical quantities can become arbitrarily large or small without introducing any contradiction. Still, it is necessary to be extremely careful whenever introducing infinities into physical systems. For if any quantity is actually infinitely large or small, then every quantity to which it is arithmetically related must also be either actually infinitely large or small. Infinities often entail physical inconsistencies.

Since the transfinite includes the finite, if U is any hyperdigital universe, then all the fundamental physical quantities in U are finitely or transfinitely recursive. Say a physical quantity is hyperdigital if and only if it is measured by some infinitely long series of finite digits. Integers and rational numbers (fractions) are hyperdigital. A real number is hyperdigital (it is a recursive real number) if and only if there is some recursive rule for generating its series of digits. For example: π is hyperdigital since there is a recursive rule for generating each digit of the infinite series 3.14159.... If U is hyperdigital, then physical quantities (e.g. mass, charge, length, time) can be infinitely precise. Arithmetical operations in U, for example, can be infinitely precise manipulations of fractions.

It is possible for space and time in hyperdigital universes to be infinitely subdivided (infinite extension raises subtle problems). One way to

define an infinitely subdivided space recursively is by endlessly many insertions of cells between cells. Recall the construction of the finite grids $G(i)$ shown in Figure 13.1. We extend the construction to the transfinite by adding a limit clause for the infinitely subdivided grid $G(\infty)$. $G(\infty)$ is the union of all the $G(i)$ for finite i. $G(\infty)$ is a *dense* 2D cellular space: between any 2 cells, there is always another cell. The topology of $G(\infty)$ is not the same as the topology of an infinitely extended 2D lattice (e.g. an infinitely large chess board). An infinitely extended 2D lattice is not dense. Even though $G(\infty)$ is dense, each cell (each point in $G(\infty)$) still has exactly 8 neighbors. Since every cell has 8 neighbors, it is possible to run rules from finite 2D cellular automata (CAs) on $G(\infty)$. If time is kept discrete, then Conway's Game of Life CA can run on $G(\infty)$. It is also possible to recursively define dense time by always inserting moments between moments on the temporal dimension. It is easy to build lattice gasses and other CAs on the dense grid. There is a large class of digital universes on dense lattices. These universes have infinitely complex dynamics.

One way to define an infinitely subdivided time is to use acceleration: each change happens twice as fast. For example: accelerating Turing machines (ATMs) are infinitely complex dynamical systems (Copeland 1998). An ATM consists of a Turing machine read/write head running over an actually infinitely long tape. An ATM tape can have infinitely many 1s, unlike a classical TM tape. An ATM is able to accelerate. If it performs an act at any speed, it can always perform the next act twice as fast. ATMs can perform *supertasks* (Koetsier & Allis 1997). An ATM starts with some initial tape-state T_0 at time 0. It computes at Zeno points. It performs the first computation in 1/2 seconds that prints T_1 at time 1/2; it performs a second computation in 1/4 seconds that prints T_2 at 3/4; it performs its nth computation in $1/2^n$ seconds to print T_n at time $(2^n - 1)/2^n$. At 1 second, the ATM has computed infinitely many operations. At the limit time 1, an ATM outputs the limit of the tape-states sequence $\{T_0, T_1, T_2, \ldots\}$, if the series converges, or a blank tape-state if it does not converge. Copeland shows that ATMs are more powerful than classical Turing machines. Programs for ATMs describe infinitely complex structural features of concrete systems and are true of universes with infinitely complex space-times (such as infinitely subdivided space-times) or infinitely complex causal regularities.

If U is hyperdigital, then its physical quantities, space-time, and causal laws are all defined by transfinite recursion. Consider a universe defined as follows: (1) the space of U is an infinitely subdivided 3D grid like $G(\infty)$; (2) the time is infinitely subdivided so that U is made up of infinitely many space-time point-instants (cells); (3) infinitely precise physical quantities (rational numbers or recursive real numbers) are associated with geometrical complexes of cells (e.g. the 0-dimensional cells in U, the 1D links between cells, the 2D areas bounded by links of cells, the 3D volumes bounded by areas of cells, etc.); (4) all the physical laws in U are transfinitely recursive functions on the rational numbers or recursive reals. Dense space-times whose causal laws are ATM transformations of infinitely long digit sequences quantities are examples of hyperdigital universes.

5 The Physics of Computation

If our universe is digital, then it is possible that some things in it are digital computers. Our universe obviously contains classical physical realizations of finite Turing machines (TMs). Therefore, it is at least finitely recursive. Since classical TMs have potentially infinitely long tapes, and can operate for potentially infinitely long periods of time, finite TMs are not really even as powerful as Turing machines. So far, all efforts to build computers more powerful than finite TMs have failed. Quantum computers do not exceed the limits of finite Turing machines (Deutsch 1985). Some suggestions for making hypercomputers involve accelerating the machinery to the speed of light or the use of unusually structured space-times. However, such suggestions are matters for science fiction. Since an ATM accelerates past any finite bound, it requires infinitely much energy to perform any infinite computation. If our universe is digital, then all the things in it are too, including human bodies and brains. If it is hyperdigital then it is possible that some things in it (some proper

parts of it) are hyperdigital computers. However, hyperdigital computers run into the lower limits imposed by quantum mechanics (e.g. the Planck time or length) or into the upper limits imposed by relativity theory (e.g. the speed of light). An accelerating Turing machine does not appear to be physically possible in our universe.

If our universe is nonrecursive, then it physically realizes properties that have neither finite nor transfinite recursive definitions. Perhaps, it physically instantiates the nonrecursive real number continuum. Analog computers are possible in universes that instantiate the nonrecursive continuum. However, the continuum is not mathematically well understood (e.g. its cardinality is undetermined; it has unmeasurable subsets; it is supposed to be well-ordered but no well-ordering is known; etc.). Attempts to define analog computation in our universe (e.g. continuously varying electrical current) conflict with the laws of quantum mechanics. If quantum mechanics is a correct description of our universe, then is unlikely that there are any analog computers in our universe. If today's biology is right, then real neural networks are not analog machines. Perhaps our universe is nonrecursive because its structure is logically undecidable. Just as Gödel's theorems prove (roughly) that there are facts about an arithmetic structure S that are not decidable within S, it may be that analogous theorems tell us that there are facts about our universe that are not decidable by any given axiomatic physical theory. The physical structure of our universe may not be axiomatizable at all. It may be deeply undecidable. Physical computation in our universe, as far as we presently know, is limited to finite Turing computability. Whatever the upper bound on physical computation in our universe, it seems clear that this bound is contingent. While hypercomputers seem both mathematically and physically possible, the internal limitations of our universe might actually prevent it from containing any.

6 Conclusion

Philosophers have long defined possible worlds as sets of propositions. Propositions are binary alternatives, either true (1) or false (0). Wheeler's slogan "Its from Bits" implies that physical reality (the "Its") is generated from binary alternatives (the "Bits"). Therefore, the "Its from Bits" program naturally hooks up with the metaphysics of possible worlds. If there are finitely many propositions in some world, and if the logical relations among them are recursive, then that world is digital. If we want to link finite digital physics to the metaphysics of possible worlds, we need to define the propositions physically. One way to do this is via Quine's suggestion of a *Democritean world* (Quine 1969: 147–52). It has been argued that Quine's theory of Democritean physics leads to a Pythagorean vision of physical reality as an entirely mathematical system of numbers. If some universe is recursive, then there is some (finitely or transfinitely) computable system of numbers that is indiscernible from it. A recursive universe is Pythagorean: physical structures are identical with numerical structures. The mystery of the link between the material and the mathematical is thereby solved: the material is reducible to the mathematical.

Since our brains and bodies are physical things, they are finitely recursive, transfinitely recursive, or nonrecursive. If human beings are somehow more than physical, then their transcendence of material reality is reflected by their computational abilities. If physics is digital, then we transcend it exactly insofar as we are hyperdigital or even nonrecursive. If physics is hyperdigital, then we transcend it exactly insofar as we are nonrecursive. Perhaps discussions of free will or our mathematical capacities aim to find the degree by which we surpass physical reality. It is possible that we are parts of hyperdigital computers even if we are only digital. The limits of our cognitive powers may be the limits of the computers that contain us, even if we are only parts of those machines, and even if those machines infinitely transcend physical computability. If physics is recursive, then there is some recursive property (a program) that is exactly instantiated by each person's body over the whole course of its life. The history or fate of each person's body is a program. If such programs exist, they are multiply realizable; so, if physics is recursive, and if all possible recursive worlds exist, then our lives (and all variations of them) are endlessly repeated

within the system of digital or hyperdigital universes. One could hardly hope for a richer kind of personal immortality. So far from eliminating the soul, recursive physics may show that it has entirely natural realizations.

References

Chopard, B. and Droz, M. 1998. *Cellular Automata Modeling of Physical Systems.* New York: Cambridge University Press. [An advanced text that discusses the use of CAs in many aspects of physical theory.]

Copeland, B. J. 1998. "Super Turing-machines." *Complexity* 4(1); 30–2. [A brief introduction to accelerating Turing machines, with many references.]

Deutsch, D. 1985. "Quantum theory, the Church–Turing principle and the universal quantum computer." *Proceedings of the Royal Society*, Series A, 400: 97–117. [The classical discussion of quantum computing.]

Finkelstein, D. 1995. "Finite physics." In R. Herken, ed., *The Universal Turing Machine: A Half-Century Survey.* New York: Springer-Verlag, pp. 323–47. [A discussion of physical theories based on the assumption that nature is finite.]

Fredkin, E. 1991. "Digital mechanics: an informational process based on reversible universal cellular automata." In H. Gutowitz, ed., *Cellular Automata: Theory and Experiment.* Cambridge, MA: MIT Press, pp. 254–70. [A discussion of the thesis that nature is finitely recursive hence a CA.]

Koetsier, T. and Allis, V. 1997. "Assaying supertasks." *Logique et Analyse* 159: 291–313. [An excellent analysis of transfinite operations.]

Quine, W. V. 1969. *Ontological Relativity and Other Essays.* New York: Columbia University Press. [Discusses Democritean worlds.]

Steinhart, E. 1998. "Digital metaphysics." In T. Bynum and J. Moor, eds., *The Digital Phoenix: How Computers are Changing Philosophy.* Oxford and Malden, MA: Blackwell, pp. 117–34. [An analysis of the thesis that nature is finitely recursive, with an extensive bibliography on finitely recursive physics.]

Toffoli, T. and Margolus, N. 1987. *Cellular Automata Machines: A New Environment for Modeling.* Cambridge, MA: MIT Press. [A classic work on the use of CAs in physical theory; deals nicely with CAs and differential equations.]

Wolf-Gladrow, D. 2000. *Lattice-gas Cellular Automata and Lattice Boltzmann Models: An Introduction.* Lecture Notes in Mathematics, vol. 1725. New York: Springer-Verlag.

Zeilinger, A. 1999. "A foundational principle for quantum mechanics." *Foundations of Physics* 29(4): 631–43. [Discusses the use of information theory as a basis for quantum mechanics; references to the "It from Bit" program.]

Zurek, W. H., ed. 1990. *Complexity, Entropy, and the Physics of Information.* SFI Studies in the Sciences of Complexity, vol. 8. Reading, MA: Addison-Wesley. [The classic work on the physics of information, with many important essays.]

Chapter 14

Cybernetics

Roberto Cordeschi

Introduction

The term *cybernetics* was first used in 1947 by Norbert Wiener with reference to the centrifugal governor that James Watt had fitted to his steam engine, and above all to Clerk Maxwell, who had subjected governors to a general mathematical treatment in 1868. Wiener used the word "governor" in the sense of the Latin corruption of the Greek term *kubernetes*, or "steersman." As a political metaphor, the idea of steersman was already present in A. M. Ampère, who in 1843 had defined cybernetics as the "art of government." Wiener defined cybernetics as the study of "control and communication in the animal and the machine" (Wiener 1948). This definition captures the original ambition of cybernetics to appear as a unified theory of the behavior of living organisms and machines, viewed as systems governed by the same physical laws.

The initial phase of cybernetics involved disciplines more or less directly related to the study of such systems, like communication and control engineering, biology, psychology, logic, and neurophysiology. Very soon, a number of attempts were made to place the concept of *control* at the focus of analysis also in other fields, such as economics, sociology, and anthropology. The original ambition of "classical" cybernetics thus seemed to involve also several human sciences, as it developed in a highly interdisciplinary approach, aimed at seeking common concepts and methods in rather different disciplines. In classical cybernetics, this ambition did not produce the desired results and new approaches had to be attempted in order to achieve them, at least partially.

In this chapter, we shall focus our attention in the first place on the specific topics and key concepts of the original program in cybernetics and their significance for some classical philosophic problems (those related to ethics are dealt with in Chapter 5, Computer Ethics, and Chapter 6, Computer-mediated Communication and Human–Computer Interaction). We shall then examine the various limitations of cybernetics. This will enable us to assess different, more recent, research programs that are either ideally related to cybernetics or that claim, more strongly, to represent an actual reappraisal of it on a completely new basis.

1 The Basic Idea behind Classical Cybernetics

The original research program of classical cybernetics was strongly interdisciplinary. The research fields within which cybernetics interacted can be grouped under three headings: engineering/

biology, philosophy/psychology, and logic/neuroscience.

1.1 Cybernetics between engineering and biology

The study of automatic control devices in machines attained full maturity around the middle of the twentieth century. The essence of automatic control resides in the capacity of a (usually electromechanical) system *S* to attain a goal-state *G* (the Greek word for goal is *telos*) set by a human operator, without the latter having to intervene any further to modify the behavior of *S* as it attains *G*. In this case, one may also speak of *closed loop* or *feedback control*. Engineers have mathematically described different types of closed loop, which have been used in both electronic and control engineering. A typical example is the positive feedback used in oscillators, or in the so-called regenerative receivers in the early radios, where part of the output signal is fed back in such a way as to increase the input signal. Of greater interest in this context is negative feedback. The behavior of a negative feedback system is governed by the continuous comparison made between the current state *C* and the state established as a reference parameter *G*, in such a way that the system uses this error information to avoid ever wandering too far from the latter. Watt's governor is an example of such a system: it maintains the speed of rotation of the driving shaft of a steam engine approximately constant in the face of load variations. It is thus capable of regulating itself automatically (*self-regulation*) without the need for any intervention by human operators once the latter have set the reference parameter (in this case *G* = the desired speed).

Devices like Watt's governor are the genuine and influential precursors of cybernetic machines. Examples of such self-regulating systems were known long before Watt's governor, as far back as the period of ancient Greece (Mayr 1970). On the contrary, the clockwork automata of the eighteenth century – such as the androids constructed by the Swiss watchmaker Pierre Jaquet-Droz and his son Henri-Louis – although astonishing in the realistic reproduction and the tiny size of their movements, cannot be correctly listed among the ancestors of cybernetic machines. These automata are merely "mechanic" and lack the fundamental self-regulating capacity typical of feedback control systems.

The study of the different feedback control mechanisms was common in Wiener's times, as was the analysis of self-regulation in living organisms. In the latter case, the existence of such systems, which may be compared to negative feedback devices, had already been described in modern physiology, in particular by Claude Bernard and Walter B. Cannon. Examples include systems that automatically maintain at a constant level body temperature, breathing, and blood pressure. In the late 1920s, Cannon referred to these systems as *homeostatic* systems.

Wiener's definition of cybernetics thus finds its initial justification in the converging of two research areas that, although having developed separately within engineering and biology, in Wiener's times seemed to share an essential core of common problems all strictly related to the study of control and information transmission processes, abstracted from mechanical or biological implementations. In 1943 Wiener, together with Arturo Rosenblueth, a physiologist and one of Cannon's pupils, and the engineer Julian Bigelow, wrote a brief article, entitled "Behavior, Purpose, and Teleology," in which the unified study of living organisms and machines, which a few years later was to suggest the term "cybernetics," was set out explicitly (Rosenblueth, Wiener, & Bigelow 1943).

1.2 Cybernetics between philosophy and psychology

In their 1943 article, Rosenblueth et al. actually provided considerably more than a comparative analysis of the self-regulating mechanisms in living organisms and machines. They supported a view that was immediately perceived as provocative by numerous philosophers and which gave rise to a very lively debate. The three authors, after summarizing the fundamental theoretical issues involved in the study of the new control devices, claimed that science could now reappraise the vocabulary of teleology, which included

such terms as purpose, goal, end, and means. According to them, teleological analyses had been "discredited" from the scientific point of view because of the Aristotelian notion of purpose as final cause. The term "final cause" suggests that the purpose is supposed to guide the behavior directed towards its attainment, despite the fact that, insofar as the purpose is a state to be attained (end state), it is a *future* state. Compared with the ordinary causal explanation, in which the cause always *precedes* the effect, the teleological explanation seems to give rise to a puzzle, that of the reversal of causal order. The hypothesis advanced by the founders of cybernetics was that the vocabulary of teleology might be revaluated by means of an objective or operational definition of its terms that allows the puzzle introduced by the notion of final cause to be avoided. In the definition they proposed, the "purpose," i.e. the final state *G* pursued by a system *S*, either natural or artificial, is the state that serves as a reference parameter for *S*, and *S*'s teleological behavior is nothing else but *S*'s behavior under negative feedback control. This was a provocative idea since psychologists and vitalist philosophers saw purposeful action as characterizing only the world of living organisms, and opposed the latter both to the world of artificial or synthetic machines and to the physical world in general (see, for instance, McDougall 1911). In fact, the new feedback machines, by interacting with the external environment, are capable of automatically changing the way they function in view of a given purpose. For philosophers concerned with a materialistic solution of the mind–body problem, cybernetics thus suggests how certain behavior regularities, usually classified as teleological to distinguish them from causal regularities in physics, may be described using purely physical laws and concepts.

As pointed out by the logical positivist philosopher Herbert Feigl, a champion of the materialist thesis of the identity between types of mental states and types of brain states, with the advent of cybernetics the concept of *teleological machine* was no longer a contradiction in terms (Feigl 1967). In addition, according to Feigl, cybernetics suggested the possibility of integrating the various levels of explanation, the mental and the physical, in view of a future neurological, and ultimately, physical microexplanation of the teleological behavior itself. Cybernetics could then provide further support for the Unitary Science proposed by logical neopositivism, according to which it was ultimately possible to hypothesize the reduction to physics of the concepts and methods of the other sciences.

Clearly, cybernetics was reproposing the idea of the organism-machine of the old mechanist tradition in a completely new context. The idea had already been implicit in Descartes who, in the *Traité de l'homme* (1664), had described the functioning of the human body in terms of hydraulic automatisms. Descartes had argued for a fundamental distinction between human beings and true automata, which represent the nonhuman animals in the living world. However, La Mettrie, in his *Homme machine* (1748), dropped Descartes's dualism and claimed that man himself is merely a machine with a particularly complex organization. Mechanistic conceptions of the living were proposed in the eighteenth century also by other authors, such as George Cabanis, while Thomas Huxley, referring back to Descartes's theory, claimed in the following century that man was nothing but an automaton possessing consciousness. The animal-automaton theory was then revived, in the interpretations of animal behavior, in terms of chains of reflexes in psychology and philosophy, between the eighteenth and the nineteenth centuries (see Fearing 1930).

It is again the new idea of a cybernetic machine capable of interacting with the environment that abated interest in the reflex-arch concept rampant in conventional neurological and psychological mechanisms. Indeed, instead of the simple stimulus–response relationship typical of the reflex arch, the interest was now focused on a circular relationship, or loop, through which the response could be fed back as the effect of the stimulus itself. Behavioristic psychologists like E. Thorndike and Clark L. Hull had already pointed out this aspect. Thorndike had explicitly formulated a trial-and-error learning Law of Effect, in which it was precisely the effect that reinforced the correct response among the many possible responses attempted at random by the organism during the learning phase. Between the 1920s and 1930s, Hull proposed an ambitious

research program, which he himself defined as a "robot approach," which foreshadowed that of cybernetics. The aim of the robot approach was the construction of machines that actually worked and hence could be viewed as mechanical models of trial-and-error learning and learning by conditioning. By constructing these models (which were actually very simple electromechanical devices), Hull set out to prove that it was useless to employ vitalist entities/concepts to attempt to account for mental life. Indeed, if a machine behaves like an organism – in Hull's view – the behavior of an organism may be accounted for by means of the same physical laws as those used to explain the machine's behavior. The reductionism underlying this thesis explains why Hull subscribed to the logical positivist hypothesis of Unitary Science.

Kenneth Craik, the Cambridge psychologist who, at the dawn of cybernetics, described several models of adaptation and learning based on different types of feedback, pointed out that Hull's position actually represented an innovation of mechanistic tradition. Unlike the supporters of mechanistic conceptions of life such as Cabanis and others, based on the man-machine metaphor, Hull had endeavored to construct learning models that, insofar as they were not generic metaphors but working implementations, allowed the hypothesis of the *mechanical* nature of this phenomenon to be tested (Craik 1943: 52). This observation by Craik on the nature of models is fundamental, as it sheds light on the simulative methodology later developed by cybernetics and the mental life sciences that followed his teachings.

1.3 Cybernetics between logic and the neurosciences

The interaction with logic and neurology is another feature of classical cybernetics. The biophysicist Nicolas Rashevsky had already made a mathematical analysis of several nervous functions. However, it was the article published in 1943 by Warren McCulloch and Walter Pitts that introduced logic into cybernetics (Anderson & Rosenfeld 1988). The article proposed a "formal" neuron, a simplified analog of a real neuron, viewed as a threshold unit, that is, functioning according to the "all-or-nothing law" (a neuron fires or does not fire according to whether the pulses it receives exceed a certain threshold or not). Neurons of this type could be interconnected to form networks, whose functioning could then be explored according to the laws of classic propositional logic. McCulloch and Pitts' article forms the basis of the development of artificial neural networks as well as computer science. John von Neumann, for example, adopted its symbolic notation in 1945 in his well-known *First Draft*, in which he described the computer architecture that was later named after him (all ordinary PCs have a von Neumann architecture).

Neurology had already suggested to psychologists laws of learning based on the assumption that the physical basis of learning is to be sought in the presence, in the central nervous system, of neurons whose reciprocal connections may be strengthened or weakened according to the stimuli received by the organism from the outside world. The tradition of connectionism, which dates back to Thorndike, was revived in the 1940s in the research carried out by Donald Hebb. Unlike the preceding connectionism, Hebb's approach supported a new interpretation of the nervous system containing reverberating neural loops. The presence of such loops in the brain tissue had been observed, among others, by the neurologist R. Lorente de Nó. This new representation of the nervous system now tended to replace that of the quasi-linear connections between stimulus and response, which were previously predominant. Within this new paradigm, Hebb formulated the learning law named after him, according to which a connection between two neurons activated at short time intervals tends to be strengthened (Hebb 1949).

After the official birth of cybernetics, neurological connectionism comes into contact with the neural networks *à la* McCulloch and Pitts in the work done by Frank Rosenblatt, the builder of one of the best-known machines of the classical cybernetics era, the Perceptron. Constructed at Cornell University in 1958, the Perceptron displays an elementary form of learning, consisting in learning to discriminate and classify visual patterns, such as letter-shaped figures (Anderson & Rosenfeld 1988). In its simplified version, the

Perceptron consists of three layers: a first layer, the analog of a retina, collects the input stimuli and is composed of several units or neurons *à la* McCulloch and Pitts, randomly connected to one or more units of the second layer, the association system. The units comprising the latter, or association units, are then connected to a third layer, the effector system, which comprises the response units. The connections among the association units are modifiable: learning actually occurs through the modification of their strength or "weight." In the first Perceptron experiment, learning was based essentially on reinforcement rules. Further experiments led to the formulation of a supervised learning rule, used by the Perceptron to modify the weighting of the connection in the case of an incorrect response, leaving it unchanged when it was correct. Other neural networks have embodied quantitative statements of the Hebb rule or its modifications.

Other learning models developed during the classical cybernetics period were the mobile robots, such as those simulating an animal learning to go through a maze. Thomas Ross invented the forerunner of this type of synthetic animal, which was influenced by Hull's robot approach. In collaboration with the behavioristic psychologist Stevenson Smith, in 1935 Ross constructed a "robot rat" at Washington University. Much more interesting as models of simple learning forms, as well as being more popular, are the robots constructed by Walter Grey Walter at the Burden Neurological Institute, in England, the electronic "tortoises." The simplest of these could successfully avoid any obstacles in their path; other, more complex ones, learned by conditioning to react to different visual and auditory stimuli (Walter 1953). The tortoises had a very simple structure. They were composed of only a small number of internal units, and Grey Walter considered this to be confirmation of the assumption that, in order to account for relatively complex behavior by organisms, it is not so much the number of neurons as the number of their connections that accounts for the relatively complex behavior of living organisms.

In the newborn field of cybernetics, again in England, William Ross Ashby was perhaps the first to investigate the physical bases of learning. As early as 1940, he described in terms of equilibration the allegedly "teleological" processes of the adaptation of organisms to the environment, anticipating the aforementioned claim of Rosenblueth, Wiener, and Bigelow. According to Ashby, trial-and-error adaptation "is in no way special to living things, . . . it is an elementary and fundamental property of matter" (Ashby 1945: 13). In order to test this hypothesis, Ashby constructed a machine that he described in his book *Design for a Brain* (Ashby 1952), as the "Homeostat," with obvious reference to Cannon. The Homeostat embodied a new and important concept, that of "ultrastability," in addition to that of feedback control or "stability." In Ashby's definition, a system is said to be "ultrastable" when it is not only capable of self-correcting its own behavior (as in the case of feedback control systems), but is also capable of changing its own internal organization in such a way as to select the response that eliminates a disturbance from the outside from among the random responses that it attempts. In this way, a system such as the Homeostat is capable of spontaneously reestablishing its own state of equilibrium: it thus displays a simple form of *self-organization*. The notion of ultrastability was deemed more interesting than that of simple stability based on feedback control because it pointed the way to simulating in an artifact some features of the plasticity and variability of response typical of animal behavior. For example, according to Ashby, ultrastability could be considered on the basis of Thorndike's Law of Effect.

2 Limits of Classical Cybernetics and New Developments

All these lines of research soon entered into crisis and were drastically curtailed, when not actually abandoned, between the 1960s and 1970s. This happened mainly because of the early successes of a new discipline, which resumed the ambition of cybernetics to strive for the unified study of organisms and machines, although on a completely different basis, namely Artificial Intelligence (AI). Subsequently, several research programs typical of cybernetics were resumed, including an explicit attempt to reformulate a

"new cybernetics." In the present and the following section, we shall examine the concepts and principal results characterizing these different phenomena and their significance for the philosophy of mind and epistemology.

2.1 Teleological machines and computer programs

The claim that purpose may be defined operationally, by means of the negative feedback notion, was challenged by many philosophers, who argued that the latter does not really fulfill all the conditions for appropriately considering a behavior pattern as purposeful. In the first instance, such a definition is always relative to the external observer who *attributes* purposes *to* the system, while it tells us nothing about the purposes *of* the system. Furthermore, in any such system it is the feedback signals from the object or the state pursued as a goal *existing* in the external environment that guides the system's purposeful behavior. In the case of non-existent objects, which may nevertheless be the content of beliefs or desires of the system, the cybernetic approach seems to have nothing to say (see, for example, Taylor 1966).

Pioneers of AI further criticized the incapacity of the artifacts proposed by cybernetics, such as neural networks or systems with simple self-organizing capability, to simulate cognitive processes. They pointed out that, in order to reproduce artificial teleological behavior in a system, such as making inferences or problem-solving, it was necessary to study selection and action-planning procedures that, in the case of an artificial system, could be realized by a computer program (see Chapter 9, The Philosophy of AI and its Critique). Actually, early AI programs were considered teleological systems, although in this case the purposes were represented as symbol structures holding information about the goals pursued. Other symbol structures were used to organize the system's behavior into complex hierarchies, such as processes for creating subgoals, selecting the methods for attempting them, testing them, and so on. Two good examples are chess playing and theorem proving, the task environments preferred by early AI. In them, the problem-solver constructs an internal representation of the problem space and works out plans aimed at finding a solution, or the final state or goal, within this space. In these cases, it is not necessary for the teleological activity to be guided by a final state that actually exists in the external environment (see Pylyshyn 1984 for further details).

As regards the simulation of cognitive processes, the introduction of the concept of algorithm, which underlies the concept of program, represented an undisputed step forward and led to the development of Cognitive Science. Prompted by the notion of algorithm, or, more precisely, of a Turing machine (see Chapter 1, Computation), is a philosophic position critical of reductionist materialism in the mind–body problem. This is functionalism, which was introduced in the philosophy of mind by Putnam in his seminal article "Minds and Machines" (1960). Putnam argued that mental states could be studied not by referring them directly to brain states, but on the basis of their functional organization, that is, of their reciprocal interactions and interactions with the sensory inputs and behavioral outputs (see Chapter 9).

2.2 Neural networks

The early success of AI in constructing computer programs that could tackle significant cognitive tasks further hindered research on neural networks, the descendants of the Perceptron. The decline in research on neural networks became generalized after the publication of Minsky & Papert 1969, which demonstrated the effective difficulties encountered by the Perceptrons in discerning even very simple visual stimuli. Despite these early failures, several researchers in different countries, such as James Anderson, Eduardo Caianiello, Stephen Grossberg, and Teuvo Kohonen, continued to work on neural networks (Anderson & Rosenfeld 1988). Rosenblatt's work was finally vindicated in the early 1980s by two events, accompanied by the development of large computers, allowing the hitherto impossible simulation of complex neural networks. John Hopfield demonstrated that symmetrical neural networks necessarily evolve towards steady states,

later called "attractors" in dynamical system theory (see Chapter 3, System: An Introduction to Systems Science), and can function as associative memories (Anderson & Rosenfeld 1988). David Rumelhart and collaborators published a series of papers based on a "parallel distributed processing" (PDP) approach to information, showing how a learning algorithm based on error correction, known as "back-propagation," made it possible to overcome the main limitations of neural networks reported by Minsky and Papert. These limitations were actually found to apply only to networks with a single associative-unit layer, such as the simple Perceptron, but not to multilayer nets, that is, networks with *more than one* layer of associative units "hidden" between the input layer and the output layer. Since the 1980s, research on neural networks differing even more substantially from the original Perceptrons has flourished, and numerous models are currently available both in neurology and psychology (see Chapter 10, Computationalism, Connectionism, and the Philosophy of Mind). This research, to the extent to which it proposes models with an architecture closer to that of the real brain than the algorithmic models proposed by AI and Cognitive Science, seems to provide strong arguments to reject functionalism. The debate has resuscitated materialist-reductionist solutions of the mind–body problem (e.g. Churchland 1986) that are reminiscent of the kind of positions we have seen above to be popular during the age of classical cybernetics.

2.3 New robotics

The construction of mobile robots such as the "tortoises" very soon came to a standstill owing to the above-mentioned predominance in AI of interest in the procedures of reasoning, planning, and problem-solving. A new kind of robot was constructed based on the idea that an agent must have an explicit symbolic representation, or centralized model of the world, in order to act in it successfully. The rather disappointing results obtained in this sector in the 1970s encouraged a different kind of approach that, although connected to cybernetics, acknowledged its limitations and tried to overcome them.

Rodney Brooks has pointed out the limits of both AI robotics and cybernetic robotics clearly. First, cybernetic robotics did not take into consideration fully the possibility of decomposing complete behavior into simpler modules with the task of controlling actions that are more elementary. Secondly, cybernetic robotics either did not recognize or else underestimated the potential of digital computation and its greater flexibility *vis-à-vis* analog computation. In conclusion, "the mechanisms of cybernetics and the mechanisms of computation were intimately interrelated in deep and self-limiting ways" (Brooks 1995: 38). The new architecture proposed by Brooks appears as a radical alternative to the AI robotics approach and at the same time represents an attempt to identify a level of abstraction that would allow the limitations of cybernetic robotics to be overcome. Brooks' "subsumption architecture" describes the agent as composed of functionally distinct control levels, a possibility ignored in cybernetic robotics. These control levels then act on the environment without being supervised by a centralized control and action planning center, as is the case instead in AI robotics. In the subsumption architecture, the low-level control routines, operating via continuous feedback loops with the environment, are connected to high-level routines that control a more complex behavior. For instance, the robot Allen, the first member of this generation of new robots or "creatures," is capable of avoiding different persons and obstacles in its path (a low-level, essentially reactive task) while continuing to pursue a goal assigned to it (that is, a higher level task). Brooks' approach and that of behavior-based robotics in general, are constrained by the fact that, in the end, it is not easy to integrate an increasing number of elementary modules to obtain behaviors that are more complex. Evolutionary robotics, based on genetic algorithms, is an attempt to get round these difficulties. In general, these approaches to robotics have several advantages, such as robustness and the capability of real-time response. However, the trade-off consists of limitations imposed on planning and reasoning capabilities.

Behavior-based robotics and evolutionary robotics have had the merit of attracting attention to the importance of several aspects neglected

by early AI and by radical functionalism, namely developmental issues in cognition and the fact that the intelligence of an agent cannot easily be "disembodied," since it is also the result of the deep interaction between the agent and its environment. This accounts for the importance acquired by "situated" cognition and for the revaluation of the role of perception and of the body in general, as well as for the attention devoted to what Steven Harnad has defined the "grounding problem," i.e. the problem of grounding the meaning of symbols in an artificial system on reality (Clark 1997).

3 Self-organization and Complexity

In the preceding section, we have discussed the limits of the cybernetics program. In doing so, we have identified several research programs that were developed in opposition to this program, as in the case of symbolic AI, or else could be ideally linked to this program, such as the new neural networks and the new robotics approaches. The latter research programs are able to overcome at least some of the limitations of early cybernetics, and do so in open opposition to symbolic AI. In the present section, we shall look at other developments that, again in opposition to symbolic AI, are explicit projects for a new cybernetics. Before doing so, however, some further developments of the original cybernetics program must be briefly sketched.

3.1 Cybernetics and the human sciences

The value of Wiener' cybernetics hypothesis was confirmed by the development of control theory and the spread of the new negative feedback mechanisms, and by the discovery of automatic regulatory processes in living organisms, comparable to those of negative feedback. This led to several attempts to develop cybernetic models of the functions of living organisms (McFarland 1971). Soon, however, a much more radical approach began to gain popularity: the basic ideas of cybernetics, i.e. feedback and information control, could be applied also to the study of a very wide range of all sorts of forms of interaction among organisms or agents. In this way, cybernetics began to be used as a meeting ground for specialists in widely differing disciplines, as is shown by the Macy Foundation Conferences held in New York between 1946 and 1953. The involvement of neurologists and psychologists proved inevitable from the outset. "He who studies the nervous system cannot forget the mind, and he who studies the mind cannot forget the nervous system," said Wiener, so that "the vocabulary of the engineers soon became contaminated with the terms of the neurophysiologists and the psychologists" (Wiener 1948: 18, 15). In addition to the presence of neurologists (e.g. Rafael Lorente de Nó) and psychologists (e.g. Kurt Lewin), those historic interdisciplinary seminars were also attended by pioneers of computer science and of information theory (e.g. Claude Shannon), as well as by sociologists (e.g. Paul Lazarsfeld), ecologists (e.g. George E. Hutchinson), and social scientists (e.g. Gregory Bateson). The negative-feedback principle soon became a universal principle by means of which to interpret the evolution towards an equilibrium state of a wide range of complex systems – social, political, pedagogical, economic, industrial, and ecological. Laws belonging to specific disciplines, such as Maupertuis's principle in physics or that of Le Châtelier in chemistry, as well as different laws describing optimization phenomena in economics and interspecies interaction in biology, were to appear as examples of this unique universal principle. Inevitably, parallel to, and often mingled with, the work of the various researchers, who were trying out new conceptual synthesis tools on specific problems, there arose a popular philosophy of cybernetics that sometimes ended up employing cybernetic concepts metaphorically, going as far as to interpret the notion of feedback as the "revealer of nature's secret." It was Wiener himself who appealed against the "excessive optimism" of all those who, like Bateson and the anthropologist Margaret Mead, believed it possible to apply the ideas of cybernetics to anthropology, sociology, and economics, overestimating "the possible homeostatic elements in the community," and ultimately

turning them into the cornerstone of an approach to complexity. More generally, while cybernetics was to suggest an extension of the natural-science method to the human sciences, in the hope of repeating in the latter field the successful results obtained in the former, to the excessive optimism was actually added a "misunderstanding of the nature of all scientific achievement" (Wiener 1948: 162). Cybernetics, which "is nothing if it is not mathematical," would end up by encouraging a fashion, already rampant according to Wiener, consisting of the inappropriate use of mathematics in the human sciences: "a flood of superficial and ill-considered work" (Wiener 1964: 88).

3.2 Systems theory and second-order cybernetics

Wiener's call for caution did not prevent others from transferring the fundamental concepts of cybernetics to wider-ranging, different interdisciplinary projects. The project for a "general system theory," initially proposed by the biologist Ludwig von Bertalanffy, is a good example. Bertalanffy, while emphasizing the interdisciplinary nature of the cybernetic approach, also argued against what he believed to be its limits. His approach was not based on a homeostatic system that can be described in terms of feedback control, but on a system that exchanges matter and energy with the environment, the only system that may be defined as thermodynamically *open*. Moreover, in its more general definition, a system is a complex of elements in dynamic interaction. Bertalanffy's idea was that the cybernetic model presupposes this more general definition insofar as the feedback occurs as a "secondary regulation." It comes into play in order to stabilize elements of the system that are *already* part of the dynamic interaction that characterizes the "primary regulation" of a thermodynamically open system such as a living organism, a social body, a biological species, an industrial organization, and so on (Bertalanffy 1968). Ilya Prigogine has further developed this approach in the study of systems far from equilibrium, and by theories studying chaotic systems and complex dynamic systems (see Chapter 3).

Other authors shift the emphasis away from the notion of control, as introduced by Wiener, on to the concepts of self-organization and autonomy. These authors are closer to Ashby, who had insisted on the centrality of these notions. They focus their attention on a classic topic in the philosophy of knowledge: the relationship between the subject, or observer, and the object, or what is observed. According to these "new cyberneticians," Wienerian cybernetics, although acknowledging that the agent and its environment must be viewed as a single system, fails to place sufficient emphasis on the autonomous or "autopoietic" nature, to use the expression coined by Humberto Maturana and Francisco Varela (1987), of this interaction. In this view, reality itself becomes an interactive object, as observer and the observed exist in a perpetually unbroken circular system. The new cyberneticians thus criticize philosophic realism, which they claim was not completely ruled out by Wienerian cybernetics, and in fact is a distinctive feature of symbolic AI because of its representational view of mind. These authors consider the activity of *knowing* not as an act of duplicating or replicating, through internal (symbolic) representations, what is supposed to be already in the outside world, but as a process built up by the observer. They want to break free from what they claim to be the scientific-philosophic "dogma" *par excellence*, that is, that the aim of science should be to approach as closely as possible a fully preconstituted reality alleged to exist as such, independently of the observer.

The criticism of these epistemological claims has its starting points in Heinz von Foerster's "second-order cybernetics" and Silvio Ceccato's "operational methodology" (Somenzi 1987). Criticisms of this kind also give rise to a reappraisal of hermeneutic positions based on the central role of interpretation and language in knowledge. The outcome is twofold. On the one hand, there is the "radical constructivism" of Ernst von Glasersfeld, according to which it is the subject *S* that constructs what *S* knows, and *S* does so on the basis of *S*'s own experience – the only "world" in which *S* is capable of living (von Glasersfeld 1995). On the other hand, there are more general worldviews (those suggested, for instance, by Winograd & Flores 1986, and

above all by Varela, Thompson, & Rosch 1991), in which situated cognition and constructivism, autopoiesis and the hermeneutics of Hans Gadamer, the philosophy of Martin Heidegger and Buddhist philosophy are occasionally gathered together in a criticism of the alleged, Western, "scientist" or "rationalist" tradition, variously defined as "Cartesian" or "Leibnizian." It is still unclear whether these positions bring any advancement in our understanding of cybernetic-related phenomena. On the other hand, many important and legitimate requirements underlying these positions seem to be already fulfilled by the very tradition that they are challenging, whenever the latter is not caricatured or oversimplified (see, for example, Vera & Simon 1993).

Acknowledgments

I am grateful to Giuseppe Trautteur and to Luciano Floridi for their comments on a previous version of this paper.

References

Anderson, J. A. and Rosenfeld, E., eds. 1988. *Neurocomputing.* Cambridge, MA: MIT Press. [This book collects classical works in the history of neural networks, from those by McCulloch & Pitts, Hebb, and Rosenblatt, to those by Anderson, Caianiello, Kohonen, Grossberg, Hopfield, Rumelhart, and others. A second volume was published in 1990.]

Ashby, W. R. 1945. "The physical origin of adaptation by trial and error." *Journal of General Psychology* 32: 13–25. [One of the clearest statement of the newborn cybernetics.]

——. 1952. *Design for a Brain.* London: Chapman and Hall; 2nd ed. New York: Wiley, 1960. [This book is usually considered a classic of cybernetics. It synthesizes Ashby's research on the physical bases of adaptation and learning and on the concept of self-organization.]

Bertalanffy, L. von 1968. *General System Theory.* New York: Braziller. [The reference work on system theory.]

Brooks, R. A. 1995. "Intelligence without reason." In L. Steels and R. Brooks, eds., *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents.* Hillsdale, NJ: Erlbaum. [This essay includes a criticism of the so-called "symbolic" view of intelligence from the viewpoint of the new robotics, of which the author is one of the main proponents.]

Churchland, P. S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* Cambridge, MA: MIT Press. [A book that includes both an introduction to the neuroscience addressed to philosophers and a criticism of the main claims of classical Cognitive Science.]

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press. [Debates the main claims of situated cognition and gives a well-balanced critical judgment.]

Craik, K. J. W. 1943. *The Nature of Explanation.* Cambridge: Cambridge University Press. [A book that foreruns several issues of both cybernetics and cognitive science.]

Fearing, F. 1930. *Reflex Action: A Study in the History of Physiological Explanation.* Cambridge, MA: MIT Press. [This book is still an excellent introduction to the historical and the epistemological issues of mechanism in neurophysiology.]

Feigl, Herbert. 1967. *The "Mental" and the "Physical": The Essay and a Postscript.* Minneapolis: University of Minnesota Press.

Hebb, D. O. 1949. *The Organization of Behavior.* New York and London: Wiley and Chapman, New York and London. [The book that is currently the reference text of new connectionists.]

Hook, S., ed. 1960. *Dimensions of Mind: A Symposium.* New York: New York University Press. [A reading including classical articles by Feigl, Putnam, McCulloch, and others that demonstrate the influence of cybernetics and Turing-machine functionalism on the debate regarding the mind–body problem.]

Maturana, H. and Varela, F. J. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding.* Boston: New Science Library. [Maturana, a pioneering cybernetician, develops with Varela the notion of "autopoietic circle."]

Mayr, O. 1970. *The Origins of Feedback Control.* Cambridge, MA: MIT Press. [A historical survey of several feedback control systems before the advent of cybernetics.]

McDougall, W. 1911. *Body and Mind: A History and a Defense of Animism.* London: Methuen. [A passionate defense of vitalism and a criticism of the different mechanistic solutions of the mind–body problem.]

McFarland, D. J. 1971. *Feedback Mechanism in Animal Behavior*. New York: Academic Press. [An approach to the life sciences based on feedback control models.]

Minsky, M. L. and Papert, S. 1969. *Perceptrons*. Cambridge, MA: MIT Press; repr. 1988, with the authors' Preface and Postfaction. [The classic critical analysis of the limitations of the early Perceptrons.]

Putnam, Hilary. 1960. "Minds and machines." In Sidney Hook, ed., *Dimensions of Mind: A Symposium*. New York: New York University Press, 1960.

Pylyshyn, Z. W. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press. [An attempt to give a foundation to classical cognitive science, as opposed to both behaviorism and connectionism.]

Rosenblueth, A., Wiener, N., and Bigelow, J. 1943. "Behavior, purpose and teleology." *Philosophy of Science* 10: 18–24. [The manifesto of the newborn cybernetics.]

Somenzi, V. 1987. "The 'Italian Operative School.'" *Methodologia* 1: 59–66. [This paper analyzes Ceccato's claims, who was, with von Foerster, one of the first critics of realism in the epistemology of cybernetics.]

Taylor, R. 1966. *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall. [A philosophical criticism of the mechanistic and cybernetic interpretations of purpose.]

Varela, F. J., Thompson, E., and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA; MIT Press. [Heidegger and Buddha against Descartes and Leibniz. A criticism of classical cognitive science, partially based on certain claims of cybernetics.]

Vera, A. H. and Simon, H. A. 1993. "Situated action: a symbolic interpretation." *Cognitive Science* 17: 7–48. [A lively response to recent criticisms of the so-called "classical paradigm" in cognitive science.]

von Glasersfeld, E. 1995. *Radical Constructivism*. Brighton, UK: Falmer Press. [The author introduces radical constructivism and examines the constructivist strand in the history of philosophy.]

Walter, W. G. 1953. *The Living Brain*. London: Duckworth. [The book includes a clear description of the famous electronic "tortoises" within the framework of the mechanistic hypotheses of cybernetics.]

Wiener, N. 1961 [1948]. *Cybernetics, or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, MA: MIT Press. [The book that made cybernetics popular, written by its founder.]

——. 1964. *God & Golem, Inc*. Cambridge, MA: MIT Press. [A clear exposition of the ideas of cybernetic, of its hopes and fears.]

Winograd, T. and Flores, F. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex. [A criticism of the "classical paradigm" of cognition, of which one of the authors (Terry Winograd) has been one of the most authoritative proponent.]

Chapter 15

Artificial Life

Mark A. Bedau

Artificial life (also known as "ALife") is a broad, interdisciplinary endeavor that studies life and life-like processes through simulation and synthesis. The goals of this activity include modeling and even creating life and life-like systems, as well as developing practical applications using intuitions and methods taken from living systems. Artificial life both illuminates traditional philosophical questions and raises new philosophical questions. Since both artificial life and philosophy investigate the essential nature of certain fundamental aspects of reality like life and adaptation, artificial life offers philosophy a new perspective on these phenomena. This chapter provides an introduction to current research in artificial life and explains its philosophical implications.

The Roots of Artificial Life

The phrase "artificial life" was coined by Christopher Langton. He envisioned a study of life as it could be in any possible setting, and he organized the first conference that explicitly recognized this field (Langton 1989). There has since been a regular series of conferences on artificial life and a number of academic journals have been launched to publish work in this new field.

Artificial life has broad intellectual roots, and shares many of its central concepts with other, older disciplines: computer science, cybernetics, biology, complex systems theory, and artificial intelligence, both symbolic and connectionist (on these topics see Chapter 3, System: An Introduction to Systems Science, Chapter 9, The Philosophy of AI and its Critique, and Chapter 14, Cybernetics).

John von Neumann (1966) implemented the first artificial-life model (without referring to it as such), with his famous creation of a self-reproducing, computation-universal entity, using cellular automata. Von Neumann was trying to understand some of the fundamental properties of living systems, such as self-reproduction and the evolution of complex adaptive structures. His approach was to construct simple formal systems that exhibited those properties. This constructive and abstract methodology typifies contemporary artificial life, and cellular automata are still widely used in the field.

At about the same time, cybernetics (Wiener 1948) applied two new tools to the study of living systems: information theory and the analysis of self-regulatory processes (homeostasis). One of the characteristics of living systems is their spontaneous self-regulation: their capacity to maintain an internal equilibrium in the face of changes in the external environment. This

capacity is still a subject of investigation in artificial life. Information theory concerns the transmission of signals independently of their physical representation. The abstract and material-independent approach of information theory is characteristic of artificial life.

Biology's contribution to artificial life includes a wealth of information about the life forms found on Earth. Artificial life seeks to understand all forms of life that could exist anywhere in the universe, and detailed information about life on Earth is one good clue about this. Biology has also provided artificial life with models that were originally devised to study a specific biological phenomenon. For example, random Boolean networks (discussed below), which were originally devised by Stuart Kauffman as models of gene regulation networks, are now a paradigm of artificial-life research.

Physics and mathematics have also had a strong influence on artificial life. One example is the study of cellular automata as exemplars of complex systems (Wolfram 1994). In addition, artificial life's methodology of studying model systems that are simple enough to have broad generality and to permit quantitative analysis was pioneered in statistical mechanics and dynamical systems. For example, the Ising model consists of a lattice of up and down "spins" that have simple local interactions and that are randomly perturbed by "thermal" fluctuations. This model is so abstract that it contains almost none of the detailed internal physical structure of such materials as a cup of water or a bar of iron. Nevertheless, the model provides a precise quantitative description of how liquid water turns into water vapor or a bar of iron loses its magnetization as temperature rises.

Artificial life also has deep roots in artificial intelligence (AI). Living and flourishing in a changing and uncertain environment seems to require at least rudimentary forms of intelligence. Thus, the subject-matter of artificial life and AI overlap. Their methodology is also similar, since both study natural phenomena by building computational models. The computational methodology of artificial life is especially close to the connectionist movement that has recently swept through AI and cognitive science.

The Methodology of Artificial Life

The computer-model methodology of artificial life has several virtues. The discipline of expressing a model in feasible computer code requires precision and clarity. It also ensures that hypothesized mechanisms are feasible. Computer models also facilitate the level of abstraction required of maximally general models of phenomena. The bottom-up architecture of artificial-life models creates an additional virtue. Allowing microlevel entities continually to affect the context of their own behavior introduces a realistic complexity that is missing from analytically studied mathematical models. Analytically solvable mathematical models can reveal little about the global effects that emerge from a web of simultaneous nonlinear interactions. The obvious way to study the effects of these interactions is to build bottom-up models and then empirically investigate their emergent global behavior through computer simulations.

There is an important difference between the modeling strategies AI and ALife typically employ. Most traditional AI models are top-down-specified serial systems involving a complicated, centralized controller that makes decisions based on access to all aspects of global state. The controller's decisions have the potential to affect directly any aspect of the whole system. On the other hand, many natural living systems exhibiting complex autonomous behavior are parallel, distributed networks of relatively simple low-level "agents" that simultaneously interact with each other. Each agent's decisions are based on information about only its own local situation, and its decisions directly affect only its own local situation. ALife's models characteristically follow nature's example. The models themselves are bottom-up-specified parallel systems of simple agents interacting locally. The models are repeatedly iterated and the resulting global behavior is observed. Such lower-level models are sometimes said to be "agent-based" or "individual-based." The whole system's behavior is represented only indirectly. It arises out of the interactions of a collection of directly represented parts ("agents" or "individuals"). Two ALife

models illustrating this pattern are described in this section below.

The parallel, distributed character of ALife models is similar to the structure of the models studied in the connectionist (parallel distributed processing, neural network) movement. Both involve bottom-up models in which a population of autonomous agents follows simple local rules. In fact, the agents in many artificial-life models are themselves controlled by internal connectionist nets. But there are at least three important differences between typical artificial-life models and the connectionist models that have attracted the most attention, such as feedforward networks that learn by the back-propagation algorithm.

- First, artificial life and connectionism depend on different kinds of learning algorithms. Connectionist models often employ supervised learning algorithms like back-propagation. These learning algorithms are typically turned on when the network is learning and then turned off when the acquired information is applied. This distinction between training and application phases is sometimes unnatural. In addition, supervised learning algorithms require an omniscient teacher, which is also often unnatural. By contrast, the learning algorithms employed in artificial-life models usually avoid these criticisms. They are typically unsupervised and in continual operation. Often the algorithm is simply natural selection.
- Second, human intervention and interpretation play different roles in artificial life and connectionism. Typical connectionist models passively receive sensory information prepackaged by a human designer and produce output that must be interpreted by a human designer. In artificial-life models, on the other hand, a microlevel agent's sensory input comes directly from the environment in which the agent lives. In many cases, this environment is itself part of the computer model. A human designer originally creates the model, of course, but the specific way it impinges on the agents is typically the result of an unpredictable collection of low-level interactions in the model. In ALife models the microlevel agents' output is to perform actions in their environment, and those actions have direct consequences for the agents' well-being. Thus their output has an intrinsic meaning regardless of human interpretation.
- Third, artificial life and connectionism typically seek different kinds of dynamical behavior. Much connectionist modeling aims to produce behavior that settles into an equilibrium. This is because both learning and applying knowledge are conceived as fixed and determinate goals. By contrast, artificial life views much of the distinctive behavior of living systems as a process of continual creative evolution, so the aim of many ALife models is an open-ended evolutionary dynamic that is forever far from equilibrium.

The biological world is often viewed as a nested hierarchy of levels. These levels include (among other things) chemicals, organelles, cells, organs, organisms, and ecologies. Artificial-life models usually explicitly represent one level with the aim of generating the characteristic phenomena of a higher level. One of the ambitious goals of artificial life is the search for a single model that generates the behavior of all these levels from the explicit specification of only the lowest level. So far, the field has had difficulty producing a model that generates even two levels of emergent phenomena.

The most primitive phenomenon explored by some artificial-life models is self-organization. Such models study how structure can emerge from unstructured ensembles of initial conditions, such as models of chemical soups in which fundamental structures such as self-maintaining autocatalytic networks might be seen to emerge. A host of models target the organismic level, sometimes with significant interactions between organisms. These models typically allow changes in the organisms as part of the system's dynamics (e.g., through a genetic mechanism). The most common goal of research using these models is to identify and elucidate structure that emerges in the ensuing evolutionary process. Some models fit in between the chemical level and the organismic level, aiming to understand development by modeling interacting cells. Other

models are interorganismic, in the sense that they aim explicitly to model interactions between different types of organisms or agents. These models often contain elements of game theory.

Many artificial-life models are designed not to represent known biological systems but to generate wholly new and extremely simple instances of life-like phenomena. The simplest example of such a system is the so-called "Game of Life," devised by the mathematician John Conway in the 1960s (Berlekamp et al. 1982). Conway's Game of Life can be thought of as a model at the physical or chemical level, embodying an extremely simple and unique form of "chemical" interactions. However, the self-organization exhibited in the Game of Life is not a representation of chemical self-organization in the real world but a wholly new instance of this phenomenon. The Game of Life is a two-state, two-dimensional cellular automaton with a trivial nearest-neighbor rule. Think of this "game" as taking place on a two-dimensional rectangular grid of cells, analogous to a huge checker-board. Time advances in discrete steps, and a cell's state at a given time is determined by the states of its 8 neighboring cells according to the following simple "birth–death" rule: a "dead" cell becomes "alive" if and only if exactly 3 neighbors are just "alive," and a "living" cell "dies" if and only if fewer than 2 or more than 3 neighbors are just "alive." From inspection of the birth–death rule, nothing particular can be discerned regarding how the whole system will behave. But when the system is simulated, a rich variety of complicated dynamics can be observed and a complex zoo of structures can be identified and classified (blinkers, gliders, glider guns, logic switching circuits, etc.). It is even possible to construct a universal Turing machine in the Game of Life, by cunningly positioning the initial configuration of living cells. In such constructions gliders perform a role of passing signals. Analyzing the computational potential of cellular automata on the basis of glider interactions has become a major research thrust.

An example of an organismic level artificial-life system is Tierra (Ray 1992). This ALife system consists of "organisms" that are actually simple, self-replicating computer programs populating an environment consisting of computer memory and consuming CPU time as a resource. A Tierran genotype consists of a string of machine code, and each Tierran creature is a token of a Tierran genotype. A simulation starts when computer memory is inoculated with a single self-replicating program, the ancestor, which is then left to self-replicate on its own. The ancestor and its descendants repeatedly replicate, until the available memory space is teeming with creatures that all share the same ancestral genotype. To create space in memory for new descendants, older creatures are continually removed from the system. Errors (mutations) sometimes occur when a creature replicates, so the population of Tierra creatures evolves by natural selection. If a mutation allows a creature to replicate faster, that genotype tends to take over the population. Over time, the ecology of Tierran genotypes becomes remarkably diverse. Quickly reproducing parasites that exploit a host's genetic code evolve, and this prompts the evolution of new creatures that resist the parasites. After millions of CPU cycles, Tierra typically contains many kinds of creatures exhibiting a variety of competitive and cooperative ecological relationships.

Computer simulation is crucial for the study of complex adaptive systems. It plays the role that observation and experiment play in more conventional science. The complex self-organizing behavior of the Game of Life would never have been discovered without simulating thousands of generations for millions of sites. Similarly, it would have been impossible to discover the emergence of complex ecological interactions in Tierra without simulating many millions of generations. Simulation of large-scale complex systems is the single most crucial development that has enabled the field of artificial life to flourish and distinguish itself from precursors such as cybernetics.

Rather than merely producing computer simulations, some artificial-life research aims to implement systems in the real world. The products of this activity are physical devices such as robots that exhibit characteristic life-like behavior. Some of these implementations are motivated by the concern to engineer practical devices that have some of the useful features of living systems, such as robustness, flexibility, and autonomy. But some of this activity is primarily theoretical, motivated by the belief that the best way to confront

the hard questions about how life occurs in the physical world is to study real physical systems. Again, there is an analogy with biological levels. The "chemical" level is represented by work on evolvable hardware, often using programmable logic arrays, which attempts to use biologically inspired adaptive processes to shape the configuration of microelectronic circuitry. The "organismic" level is represented by new directions in biologically inspired robotics, such as using evolutionary algorithms to automate the design of robotic controllers. A swarm of robots communicating locally to achieve some collective goal is an example at the "population" level. An "ecological" level might be represented by the internet along with its interactions with all its users on computers distributed around the world.

Emergence

Both living systems and artificial-life models are commonly said to exhibit emergent phenomena; indeed, many consider emergence to be a defining feature of life. However, the notion of emergence remains ill defined. In general, emergent phenomena share two broad hallmarks: they are constituted by and generated from underlying phenomena, and yet they are also autonomous from those underlying phenomena. There are abundant examples of apparent emergent phenomena, and most involve life or mind. Yet the two hallmarks of emergence seem inconsistent or metaphysically illegitimate: How can something be autonomous from underlying phenomena if it is constituted by and generated from them? This is the problem of emergence. A solution would both dissolve the appearance of illegitimate metaphysics and enfold emergence in constructive scientific explanations of phenomena involving life and mind.

One can distinguish emergent properties, emergent entities, and emergent phenomena. Being alive, for example, is an emergent property, an organism is an emergent entity, and the life history of an organism is an emergent phenomenon. An entity with an emergent property is an emergent entity, and an emergent phenomenon involves an emergent entity possessing an emergent property. So the first step toward solving the problem of emergence is to explain the notion of an emergent property. There are three main views of what an emergent property is.

According to the first view, emergent properties apply only to "wholes" or "totalities," not to their component "parts" considered in isolation (e.g., Harré 1985, Baas 1994). For example, the constituent molecules in a cup of water, considered individually, do not have properties like fluidity or transparency, though these properties do apply to the whole cup of water. The "wholes" at one level of analysis are sometimes "parts" of a larger "whole" at a higher level of analysis, so a hierarchy can contain successive levels of this sort of emergence. This view easily explains the two hallmarks of emergence. Macrolevel emergent phenomena are constituted from and generated by microlevel phenomena in the trivial sense that wholes are constituted and generated by their constituents; and emergent phenomena are autonomous from underlying phenomena in the straightforward sense that emergent properties do not apply to the underlying entities. This notion of emergence is very broad, applies to a large number of intuitive examples of emergent phenomena, and corresponds to the compelling picture of reality consisting of autonomous levels of phenomena. Its breadth is its greatest weakness, however, for it applies to all macroproperties that are not possessed by microentities. Macroproperties are usually classified into two kinds: genuine emergent properties and mere "resultant" properties. Resultant properties are those that can be predicted and explained from the properties of the components. For example, a circle consists of a collection of points, and the individual points have no shape. So being a circle is a property of a "whole" but not its constituent "parts." Thus being a circle is an emergent property according to the first view. However, if you know that all the points in a geometrical figure are equidistant from a given point, then you can conclude that the figure is a circle. So being a circle is a resultant property. To distinguish emergent from resultant properties one must turn to other views.

The second main view construes emergent properties as supervenient properties with causal powers that are irreducible to the causal powers

of microlevel constituents (e.g., Kim 1999). On this view, supervenience explains the sense in which the underlying processes constitute and generate the emergent phenomena, and irreducible causal powers explain the sense in which they are autonomous from underlying phenomena. These irreducible causal powers give emergent properties a dramatic form of ontological novelty that many people associate with the most puzzling kinds of emergent phenomena, such as consciousness. However, an irreducible but supervenient causal power by definition cannot be explained in terms of the aggregation of the microlevel potentialities. No evident mechanism explains these irreducible supervenient powers, so they must be viewed as primitive or "brute" facts of nature. In addition, this strong form of emergence seems to be scientifically irrelevant. Illustrations of it in recent scientific literature almost universally focus on one isolated example: Sperry's explanation of consciousness from over 30 years ago (Sperry 1969). There is little if any evidence that this form of emergence is empirically relevant in the sciences studying emergent phenomena.

A third view of emergence is poised midway between the first two. It refers to the resultant aggregate global behavior of complex systems. In this sense, a system's macrostate is emergent just in case it can be derived from the system's boundary conditions and its microlevel dynamical process but only through the process of iterating and aggregating all the microlevel effects (e.g., Bedau 1997a). In this case, the microlevel phenomena clearly constitute and generate the macrolevel phenomena. At the same time, the macrolevel phenomena are autonomous in that the only way to recognize or predict them is by empirically observing the macrolevel effect of aggregating all the microlevel phenomena. In effect, this view identifies emergent properties with a special subset of resultant properties: those that cannot be predicted or explained except by empirically aggregating the interactions among microlevel entities. This form of emergence is common in complex systems found in nature. Artificial life's models also exhibit it, since their bottom-up behavior consists of the continual iteration of microlevel interactions. This view attributes the unpredictability and unexplainability of emergent phenomena to the complex consequences of myriad, nonlinear, and context-dependent local microlevel interactions. Emergent phenomena can have causal powers on this view, but only by means of aggregating microlevel causal powers. There is nothing inconsistent or metaphysically illegitimate about underlying processes constituting and generating phenomena by iteration and aggregation. Furthermore, this form of emergence is prominent in scientific accounts of exactly the natural phenomena like life and mind that apparently involve emergence. However, this form of emergence sheds no light on those mysterious emergent phenomena, like consciousness, that science still cannot explain. In addition, the autonomy of these kinds of emergent phenomena seems to be merely epistemological rather than ontological. Emergent phenomena are epistemologically autonomous in the sense that knowledge of the underlying phenomena does not provide knowledge about the emergent phenomena. However, metaphysically, the emergent phenomena seem wholly dependent on the constituent phenomena, since emergent causal powers result from microlevel causal powers. This will not satisfy those who think emergent phenomena have a strong form of ontological autonomy.

Artificial life can be expected to play an active role in the future philosophical debate about emergence and related notions like supervenience, reduction, complexity, and hierarchy. Living systems are one of the primary sources of emergent phenomena, and artificial life's bottom-up models generate impressive macrolevel phenomena wholly out of microlevel interactions. Exploration and modification of these models is a constructive way to analyze the nature and causes of different kinds of emergent phenomena.

Adaptationism

Adaptive evolutionary explanations are familiar from high-school biology. It is a cliché to explain the giraffe's long neck as an adaptation for browsing among the tops of trees, on the grounds that natural selection favored longer-necked giraffes over their shorter-necked cousins. But the

scientific legitimacy of these adaptive explanations is controversial, largely because of a classic paper by Stephen Jay Gould and Richard Lewontin (1979). Gould and Lewontin directly challenge *adaptationism*: the thesis that the activity of pursuing adaptive explanations of biological traits is a legitimate part of empirical science. They accept that adaptive explanations are appropriate in some contexts, but they despair of identifying those contexts in any principled and rigorous way. Biology provides many alternatives to adaptive explanations, such as explanations appealing to allometry, genetic drift, developmental constraints, genetic linkage, epistasis, and pleiotropy. But Gould and Lewontin complain that those alternatives receive only lip-service. The presupposition that a trait is an adaptation and so deserves an adaptive explanation is treated as untestable. The fundamental challenge for adaptationism raised by Gould and Lewontin, then, is to find some empirical method for testing when an adaptive explanation is needed. This problem is often especially acute in artificial life. Those studying artificial models have the luxury of being able to collect virtually complete data, but this mass of information only compounds the problem of identifying which evolutionary changes are adaptations.

The canonical response to Gould and Lewontin makes two claims. The first claim is that *specific* adaptive hypotheses, hypotheses about the specific nature of a character's adaptation, are testable. Second, although the *general* hypothesis that a trait is an adaptation might itself not be testable, it is a working hypothesis and empirical science normally treats working hypotheses as untestable. For example, Richard Dawkins claims that "hypotheses about adaptation have shown themselves in practice, over and over again, to be easily testable, by ordinary, mundane methods of science" (Dawkins 1983: 360ff). Dawkins's point is that specific adaptive hypotheses have observable consequences that can be checked. The canonical response reflects and explains evolutionary biology's emphasis on formulating and testing specific adaptive hypotheses. But this response does not address the fundamental challenge to adaptationism, for that challenge is about the testability of *general* adaptive hypotheses, hypotheses to the effect that a trait is an adaptation. Different specific adaptive hypotheses usually have different observable consequences. A general adaptive hypothesis entails that some specific adaptive hypothesis is true, but it gives no indication which one is true. So the general adaptive hypothesis makes no particular empirical prediction. Dawkins admits that general adaptive hypotheses cannot be tested. "It is true that the one hypothesis that we shall never test is the hypothesis of no adaptive function at all, but only because that is the one hypothesis in this whole area that really *is* untestable" (1983: 361). Dawkins can defend the appeal to adaptive explanations when a specific adaptive hypothesis has been corroborated. But in the absence of this – which is the typical situation – Dawkins must concede Gould's and Lewontin's fundamental challenge.

Artificial life has been used to develop and illustrate a new defense of adaptationism. It is argued that it is possible to test general adaptive hypotheses empirically, by recording and analyzing so-called "evolutionary activity" information collected from the evolving system (Bedau 1996, Bedau & Brown 1999). The fundamental intuition behind this method is that we can detect whether an item (gene, gene complex, genotype, etc.) is an adaptation by observing the extent to which it persists in the face of selection pressures. Whenever an item that is subject to heritable variation is "active" or expressed, natural selection has an opportunity to provide feedback about its adaptive value, its costs and benefits. If it persists and spreads through a population when it is repeatedly active, and especially if it exhibits significantly more activity than one would expect to see if it had no adaptive value, then we have positive evidence that the item is persisting *because of* its adaptive value. This means that we have positive evidence that it is an adaptation and deserves an adaptive explanation, even if we have no idea about its specific adaptive function. Since natural selection is not instantaneous, maladaptive items persist for a while before they are driven out by natural selection. Adaptations are distinguished by accruing much more activity than would be expected in a nonadaptive item. A general way to measure the activity expected of nonadaptive items is to construct a "neutral shadow" of the target system – that is, a system

that is similar to the target in all relevant respects *except* that none of the items in it have any adaptive significance. The activity in the neutral shadow is a no-adaptation null hypothesis for the target system. If the target system shows significantly more activity than the neutral shadow, this excess activity must be due to natural selection and the target system must contain adaptations. The evolutionary-activity method responds directly to Gould and Lewontin. It provides an empirical method for determining when evolution is creating adaptations. Rather than just assuming that traits are adaptations, it puts this assumption to the empirical test. Another advantage of the activity method is that statistics based on activity information can be used to measure various aspects of the dynamics of adaptive evolution, thus allowing the process of adaptation in different systems to be classified and quantitatively compared (Bedau et al. 1997, Bedau et al. 1998). One weakness of the evolutionary activity method is that practical problems sometimes make activity data difficult to collect. Another weakness is that genetic hitchhikers – nonadaptive or maladaptive traits that persist because of a genetic connection to an adaptive trait – can accumulate more activity than expected in a neutral shadow. Thus, a trait that is not an adaptation can have significant excess activity if it is connected to a trait that is an adaptation. Significant excess activity in a cluster of traits shows that there are adaptations in the cluster, but it does not separate out the hitchhikers.

The adaptationist perspective on evolution emphasizes natural selection's role in creating the complex adaptive structures found in living systems. Artificial life has been the source of a new and fundamental challenge to this whole perspective. Stuart Kauffman (1993, 1995) has used artificial-life models to show that many features of metabolisms, genetic networks, immune systems, and ecological communities should be viewed not as the products of selection but largely as the spontaneous, self-organized behaviors of certain abstract complex systems. Kauffman also argues that spontaneous self-organized structures – what he calls "order for free" (Kauffman 1995) – explain both life's origin and its subsequent ability to evolve. Kauffman can make sweeping claims about order for free because the artificial-life models he studies are abstract enough to apply to a wide variety of contexts. Random Boolean networks are one such class of models. These consist of a finite collection of binary (ON, OFF) variables with randomly chosen input and output connections. The state of each variable at each step in discrete time is governed by some logical or Boolean function (AND, OR, etc.) of the states of variables that provide input to it. The network is started by randomly assigning states to each variable, and then the connections and functions in the network determine the successive state of each variable. Since the network is finite, it eventually reaches a state it has previously encountered, and from then on the network will forever repeat the same cycle of states. Different network states can end up in the same state cycle, so a state cycle is called an attractor. Kauffman found that the number of variables in the network, the number of connections between the variables, and the character of the Boolean functions determine many biologically crucial properties of the networks. These properties include the number and length of attractors, the stability of attractors to perturbation and mutation, etc. If the variables are highly connected, then the network's attractors contain so many states that the time it takes to traverse the attractor vastly exceeds the lifetime of the entire universe. Furthermore, any perturbation or mutation in the network causes a vast change in its behavior. For all practical purposes, the network behaves chaotically. The network acts differently when each variable takes input from only a biologically plausible number of other variables and when the variables are governed by biologically realistic Boolean functions. In this case, the network has a tiny number of attractors, it maintains homeostatic stability when perturbed, and mutations have limited consequences; in other words it exhibits "order for free." Furthermore, these biologically realistic Boolean networks explain a number of empirically observed features of biological systems, such as how the number of different cell types and cell replication times vary as a function of the number of genes per cell. Kauffman's nonadaptationist explanations of the origins of order are controversial, partly because of the sweeping scope of his analysis. But the suggestion that self-organization rather

than natural selection can explain much of the structure in living systems is plausible. The issue is not whether self-organization explains structure, but how much.

The problem of adaptationism is as acute in artificial life as it is in biology. Artificial life can make a distinctive contribution to the debate, for the evolutionary processes studied by artificial life provide many diverse examples of the process of adaptation. Furthermore, the systems can be analyzed with the kind of detail and rigor that is simply impossible to achieve in the biosphere, because the historical data are unavailable or impractical to examine. For analogous reasons, we can expect artificial life to contribute to our understanding of many other fundamental issues in the philosophy of biology, such as the nature of functions, the nature of species, whether and how selection operates at different biological levels, the nature of the niche, and the nature of the relationship between organisms and their environment.

Evolutionary Progress

The evolution of life shows a remarkable growth in complexity. Simple prokaryotic one-celled life led to more complex eukaryotic single-celled life, which then led to multicellular life, then to large-bodied vertebrate creatures with complex sensory processing capacities, and ultimately to highly intelligent creatures that use language and develop sophisticated technology. This illustration of evolution's creative potential has led some to propose a ladder-of-complexity hypothesis according to which open-ended evolutionary processes have an inherent, law-like tendency to create creatures with increasingly complicated adaptive structure. But the evolution of life is equally consistent with the denial of the ladder of complexity. The observed progression could be a contingent result of evolution rather than a reflection of any inherent tendency. The ladder-of-complexity hypothesis is difficult to test because we do not have a variety of different histories of life to compare. A sample size of one makes it difficult to distinguish inherent trends from artifacts.

Stephen Jay Gould (1989) devised an ideal way to address this issue, namely the thought experiment of replaying the tape of life. Imagine that the process of evolution left a record on a tape. Gould's thought experiment consists in rewinding the evolutionary process backward in time and then replaying it again forward in time but allowing different accidents, different contingencies to reshape the evolution of life. The evolution of life is rife with contingencies. Repeatedly replaying the tape of life with novel contingencies could produce as large a sample of evolutionary histories as desired. It would be relatively straightforward to determine whether a general pattern emerges when all the evolutionary trajectories are compared.

There is substantial controversy about the outcome of Gould's thought experiment. Gould himself suggests that "any replay of the tape would lead evolution down a pathway radically different from the road actually taken" (1989: 51). He concludes that the contingency of evolution will debar general laws like the hypothesized ladder of complexity. Daniel Dennett (1995) draws exactly the opposite conclusion. Dennett argues that certain complex features like sophisticated sensory processing provide a distinct adaptive advantage. Thus, natural selection will almost inevitably discover significantly advantageous features that are accessible from multiple evolutionary pathways. Examples of multiple independent evolutionary convergence, such as flight and eyesight, illustrate this argument. Dennett concludes that replaying life's tape will almost inevitably produce highly intelligent creatures that use language and develop sophisticated technology.

Artificial life can make a number of contributions to this debate. Experience in artificial life has shown time and again that expectations about the outcome of thought experiments like replaying life's tape are highly fallible. The only sure way to determine what to expect is to create the relevant model and observe the results of repeated simulation. In fact, artificial life is exactly where this sort of modeling activity occurs. A central goal of artificial life is to discover the inherent trends in evolving systems by devising a model of open-ended evolution, repeatedly replaying life's tape with different historical

contingencies and searching for patterns that hold across all the results. The best evidence in favor of the ladder-of-complexity hypothesis would come from showing that a tendency toward increasing adaptive complexity is the norm in such ALife models. However, no one has yet conducted the experiment of replaying life's tape, because no one has yet been able to create a system that exhibits continual open-ended evolution of adaptive complexity. Achieving this goal is one of the key open problems in artificial life (Bedau et al. 2000). All conjectures about the ladder of complexity will remain unsettled until one can actually replay the tape of life.

The Nature of Life

Philosophy traditionally addressed the nature of life but most philosophers ignore the issue today, perhaps because it seems too "scientific." At the same time, most biologists also ignore the issue, perhaps because it seems too "philosophical." The advent of artificial life raises the question anew, for two reasons. Modeling the fundamental features of living systems presupposes an understanding of life, and new artificial-life systems push the boundaries of what life could be.

There are three prominent views about the nature of life: life as a cluster of properties, life as metabolization, and life as evolution. The cluster conception takes two forms, depending on whether the properties in the cluster are taken to be individually necessary and jointly sufficient for life. Skeptics argue that life is characterized merely by a loose cluster of properties typically but not necessarily possessed by living entities. This view treats something as alive if it possesses a sufficient number of properties in the cluster, but no precise number of properties is sufficient. On this view, the diversity of living forms have only a family resemblance. Viewing life as a loose cluster of properties provides a natural explanation of why life has vague boundaries and borderline cases. Life is also sometimes characterized by a list of properties intended to provide something much closer to individually necessary and jointly sufficient conditions. Ernst Mayr (1982) produced a comprehensive list of such properties:

1 Living systems have an enormously complex and adaptive organization.
2 Organisms are composed of a chemically unique set of macromolecules.
3 Living phenomena are predominantly qualitative, not quantitative.
4 Living systems consist of highly variable groups of unique individuals.
5 Organisms engage in purposeful activities by means of evolved genetic programs.
6 Classes of organisms have historical connections of common descent.
7 Organisms are the product of natural selection.
8 Biological processes are especially unpredictable.

Cluster conceptions of life account for the characteristic hallmarks of life, although they do this merely by *fiat*. Lists like Mayr's raise rather than answer the question why this striking collection of features is present in an indefinite diversity of natural phenomena. The main drawback of all cluster conceptions is that they inevitably make life seem rather arbitrary or mysterious. A cluster conception cannot explain why any particular cluster of properties is a fundamental and ubiquitous natural phenomenon.

Schrödinger illustrated the second view of life when he proposed persistence in the face of the second law of thermodynamics by means of the process of metabolization as the defining feature of life.

> It is by avoiding the rapid decay into the inert state of "equilibrium" that an organism appears so enigmatic; . . . How does the living organism avoid decay? The obvious answer is: By eating, drinking, breathing and (in the case of plants) assimilating. The technical term is metabolism. (Schrödinger 1969: 75)

Living systems need some way to self-maintain their complex internal structure. So metabolization seems to be at least a necessary condition of all physical forms of life. The view that life centrally involves the process of metabolization also nicely explains our intuition that a crystal is not alive. There is a metabolic flux of molecules only at the crystal's edge, not inside it. One drawback

of metabolization as an all-encompassing conception of life is that many metabolizing entities seem not to be alive and not to involve life in any way. Standard examples include a candle flame, a vortex, and a convection cell. A second problem is whether metabolization can explain the hallmarks of life (recall Mayr's list). It is doubtful whether metabolization can explain those characteristics on Mayr's list that depend on evolution.

The third main conception of life focuses on the evolutionary process of adaptation. The central idea is that what is distinctive of life is the way in which adaptive evolution automatically fashions new and intelligent strategies for surviving and flourishing as local environments change. As John Maynard Smith explains:

> We shall regard as alive any population of entities which has the properties of multiplication, heredity and variation. The justification for this definition is as follows: any population with these properties will evolve by natural selection so as to become better adapted to its environment. Given time, any degree of adaptive complexity can be generated by natural selection. (Maynard Smith 1975: 96ff)

The view of life as evolution has two forms. Maynard Smith illustrates one form, according to which living systems are the entities in an evolving population. Recently, Bedau (1996, 1998) has argued that, in fact, an evolving system itself should be viewed as alive in the primary sense. One virtue of the conception of life as evolution is that it explains why Mayr's hallmarks of life coexist in nature. We would expect life to involve the operation of natural selection producing complex adaptive organization in historically connected organisms with evolved genetic programs. The random variation and historical contingency in the evolutionary process explains why living phenomena are especially qualitative and unpredictable and involve unique and variable individuals with frozen accidents like chemically unique macromolecules. This view can also explain why metabolism is so important in living systems, for a metabolism is a physically necessary prerequisite in any system that can sustain itself long enough to adapt and evolve. There are two main objections to viewing life as evolution. The first is that it seems to be entirely contingent that life forms were produced by an evolutionary process. The Biblical story of Adam and Eve shows that is easy to imagine life forms in the absence of any evolutionary process. A second objection calls attention to evolving systems that seem devoid of life. Viruses and prions evolve but are questionably alive, and cultural evolution provides much starker counterexamples.

The advent of artificial life has revitalized investigation into the nature of life. This is partly because one can simulate or synthesize living systems only if one has some idea what life essentially is. Artificial life's self-conscious aim to discern the general nature of life as it could be encourages liberal experimentation with novel life-like organizations and processes. Thus, artificial life both fosters a broad perspective on life and has the potential to create radically new forms of life. In the final analysis, the nature of life will be settled by whatever provides the best explanation of the rich range of natural phenomena that seem to characterize living systems. Better understanding of how to explain these phenomena will also help resolve a cluster of puzzles about life. These puzzles include whether life admits of degrees, how the notion of life applies at different levels in the biological hierarchy, whether life is essentially connected with mental capacities, and the relationship between the material embodiment of life and the dynamical processes in those materials.

Strong Artificial Life

Artificial life naturally raises the question whether artificial constructions could ever literally be alive. Agreement about the nature of life would make this question easier to answer. For example, if the defining property of living systems were the process of sustaining a complex internal organization through a metabolism, then the issue would be whether an artificially created system could literally exhibit this property (see Boden 1999 for discussion). But the debate over creating real but artificial life currently proceeds in the absence of agreement about what life is.

It is important to distinguish two questions about creating artificial life. The first concerns whether it is possible to create a physical device such as a robot that is literally alive. Aside from controversy about what life is, the challenge here is less philosophical than scientific. It concerns our ability to synthesize the appropriate materials and processes. The philosophically controversial question is whether the processes or entities inside a computer that is running an artificial-life model could ever literally be alive. This is the issue of whether so-called "strong" artificial life is possible. Strong ALife is contrasted with "weak" ALife, the uncontroversial thesis that computer models are useful for understanding living systems.

The strong ALife question is sometimes put in terms of computer simulations: can a computer simulation of a living system ever literally be alive? This formulation prompts the response (e.g., Pattee 1989, Harnad 1994) that it is a simple category mistake to confuse a *simulation* of something with a *realization* of it. A flight simulation for an airplane, no matter how detailed and realistic, does not really fly. A simulation of a hurricane does not create real rain driven by real gale-force winds. Similarly, a computer simulation of a living system produces merely a symbolic representation of the living system. The intrinsic ontological status of this symbolic representation is nothing more than certain electronic states inside the computer (e.g., patterns of high and low voltages), and this constellation of electronic states is no more alive than is a series of English sentences describing an organism. It seems alive only when it is given an appropriate interpretation. This interpretation might be fostered if the description dynamically reflects how the living system changes over time and if the simulation produces a vivid life-like visualization, but it is still only an interpretation.

A number of considerations can blunt this charge of category mistake. It is important to recognize that an artificial-life model that is actually running on a computer consists of a real physical process occurring in a real physical medium consuming real physical resources. The software specifying the model might be a static abstract entity with the ontological nature of a Platonic universal, but an actual simulation of the model has the ontological status of any physical process. Furthermore, as emphasized earlier, artificial-life models are often intended not as simulations or models of some real-world living system but as novel examples of living systems. Conway's Game of Life (Berlekamp et al. 1982), for example, is not a simulation or model of any real biochemical system. Rather, it is a simple system that exhibits spontaneous macroscopic self-organization. Similarly, Ray's Tierra (Ray 1992) is not a simulation or model of the ecology and evolution of some real biological system. Instead, it is an instance of ecological and evolutionary dynamics in a digital domain. So, when the Game of Life and Tierra are actually running in computers, they are new physical instances of self-organization and evolution. Processes like self-organization and evolution are multiply realizable and can be embodied in a wide variety of different media, including the physical media of suitably programmed computers. So, to the extent that the essential properties of living systems involve processes like self-organization and evolution, suitably programmed computers will actually be novel realizations of life. Models that merely represent some phenomenon differ from models that actually generate it. For example, a two-dimensional model of a branching process with random pruning can be viewed as a description of the evolution of more or less complex insects, if one dimension is taken to represent time and the other is taken to represent complexity. But exactly the same branching process can equally be viewed as a description of the evolution of more or less tall humans. It can even be viewed as a description of various nontemporal and nonbiological processes, such as the pattern of tributaries in a geography. In itself, the model does not intrinsically involve any of these things. By contrast, a glider in Conway's Game of Life is not an electronic pattern that is merely interpretable as a self-sustaining dynamic collective. It really *is* an electronic self-sustaining collective, whether or not anyone notices it and regards it as such. Likewise, the self-replicating machine-language programs in Ray's Tierra genuinely evolve by natural selection and genuinely engage in host/parasite relations. The nature of ALife's key problem of modeling the open-ended evolution of adaptive complexity can be appreciated in this light. It is easy to make a model that

can be interpreted as exhibiting this phenomenon; the challenge is to make a model that actually generates it.

The Turing test in artificial intelligence was an attempt to settle whether computing could be indistinguishable from thinking in the absence of any agreement about the nature of thinking itself. Thus the proposal to settle the strong ALife debate with a "Turing test" for life often arises in artificial life. Some (e.g., Sober 1992) warn that the Turing test in AI is an insufficient test for intelligence because it is possible in principle for an unthinking device to pass the test. A typical example of such a hypothetical device is a machine that stores an appropriate output for all the different input that might be encountered. The characteristic drawback of such devices is that, even to exhibit modest capabilities, the number of pieces of information they must store is larger than the number of elementary particles in the entire universe. Though possible in principle, such a device is clearly impossible in practice. Artificial life's computational methodology demands models that actually produce the phenomenon of interest. In this context, what is possible in principle but impossible in practice is irrelevant. So the experience in ALife prompts one to ignore unfeasible counterexamples to Turing tests. Harnad (1994) has advocated ecological and evolutionary indistinguishability from biological life as a Turing test for life. The motivation for this test for life is that it would be arbitrary to deny life to anything that is indistinguishable ecologically and evolutionarily from biological life. But this test is biased against life forms that are isolated from the biosphere. Systems existing inside computers running artificial-life models might exhibit all the ecological and evolutionary richness found in the biosphere. Yet they might not interact with biological life, so they might fail Harnad's test for life. Thus, Harnad's test begs the question against some forms of artificial life.

The debate about strong artificial life is intertwined with philosophical questions about functionalism and computation. A significant source of support for strong ALife is the belief that life concerns form more than matter. Although certain carbon-based macromolecules play a crucial role in the vital processes of all known living entities, metabolization creates a continual flux of molecules through living systems. Thus, life seems more like a kind of a process than a kind of material entity. This implies that life could be realized in a variety of media, perhaps including suitably programmed computer hardware. This motivation for strong ALife prompts a functionalist and computationalist view of life, analogous to contemporary functionalism and computationalism with respect to mind. Sober (1992) points out that many essential properties of organisms involve their interaction with the environment. Thus, the computational character of the processes inside organisms would not alone support functionalism and computationalism about life. But since many artificial-life models situate artificial organisms in an artificial environment, artificial life still promotes functionalism and computationalism. Bedau (1997b) argues that artificial life's models generate macrolevel dynamics with a suppleness that is distinctive of adaptive intelligence and that cannot be captured by any fixed algorithm. The models are implemented in a computer but adaptive processes like natural selection continually change the microlevel rules that govern the system. Thus, the macrolevel processes that emerge are noncomputational. This perspective still supports functionalism with respect to life, but a form of functionalism divorced from computationalism.

Artificial-life models generate behavior that is characteristic of living systems, so the practice of artificial life will continually raise the question whether a computer model of life could literally be alive. By continually challenging the boundaries between life and nonlife, artificial life will also spur novel perspectives on the issue. The debate about strong ALife will also enliven and inform many related issues in the philosophy of mind and artificial intelligence, including functionalism, computationalism, intelligence, intentionality, and representationalism.

Philosophical Methodology

Artificial life also has implications for the methodology of philosophy. Philosophy and artificial life are natural partners. Both seek to understand

phenomena at a level of generality that is sufficiently deep to ignore contingencies and reveal essential natures. In addition, artificial life's computational methodology is a direct and natural extension of philosophy's traditional methodology of *a priori* thought experiments. In the attempt to capture the simple essence of vital processes, artificial-life models abstract away as many details of natural living as possible. These models are for exploring the consequences of certain simple ideas or premises. They are "thought experiments" explored with the help of a computer. Like the traditional armchair thought experiments employed in philosophy, artificial-life simulations attempt to answer "What if X?" questions. Artificial life's thought experiments are distinctive in that they can be explored only by computer simulation; armchair analysis is simply inconclusive. Synthesizing thought experiments on a computer can bring a new clarity and constructive evidence to bear in philosophy (see Chapter 26, Computational Modeling as a Philosophical Methodology).

References

Baas, N. A. 1994. "Emergence, hierarchies, and hyperstructures." In C. G. Langton, ed., *Artificial Life III*. Redwood City, CA: Addison-Wesley, pp. 515–37. [A mathematical and technical presentation and illustration of the view of emergent properties as novel macroproperties.]

Bedau, M. A. 1996. "The nature of life." In M. Boden, ed., *The Philosophy of Artificial Life*. Oxford: Oxford University Press, pp. 332–57. [A defense of the view of life as supple adaptation or open-ended evolution, illustrated in artificial life models. For postsecondary school audiences.]

——. 1997a. "Weak emergence." *Philosophical Perspectives* 11: 375–99. [A defense of emergence as complicated iteration and aggregation of microlevel interactions, illustrated in artificial life models. For postsecondary school audiences.]

——. 1997b. "Emergent models of supple dynamics in life and mind." *Brain and Cognition* 34: 5–27. [Describes a characteristic suppleness of the dynamics of mental states, argues that artificial life models capture this kind of dynamics, and draws out the implications for functionalism about life. For postsecondary school audiences.]

——. 1998. "Four puzzles about life." *Artificial Life* 4: 125–40. [An explanation of how the view of life as supple adaptation (Bedau 1996) explains four puzzles about life. For postsecondary school audiences.]

—— and Brown, C. T. 1999. "Visualizing evolutionary activity of genotypes." *Artificial Life* 5: 17–35. [Shows how evolutionary activity graphs reveal the dynamics of adaptive evolution of genotypes in an artificial life model. For postsecondary school audiences.]

——, McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., Ikegami, T., Kaneko, K., and Ray, T. S. 2000. "Open problems in artificial life." *Artificial Life* 6: 363–76. [Describes 14 grand challenges in artificial life, each of which requires a major advance on a fundamental issue for its solution. Intelligible to secondary school audiences.]

——, Snyder, E., Brown, C. T., and Packard, N. H. 1997. "A comparison of evolutionary activity in artificial evolving systems and the biosphere." In P. Husbands and I. Harvey, eds., *Proceedings of the Fourth European Conference on Artificial Life, ECAL97*. Cambridge, MA: MIT Press, pp. 125–34. [Comparison of evolutionary activity in two artificial life models and in the fossil record reveals qualitative differences. For postgraduate audiences.]

——, ——, and Packard, N. H. 1998. "A classification of long-term evolutionary dynamics." In C. Adami, R. Belew, H. Kitano, and C. Taylor, eds., *Artificial Life VI*. Cambridge, MA: MIT Press, pp. 228–37. [Evolutionary activity is used to classify qualitatively different kinds of evolving systems. For postgraduate audiences.]

Berlekamp, E. R., Conway, J. H., and Guy, R. K. 1982. *Winning Ways for your Mathematical Plays*, vol. 2: *Games in Particular*. New York: Academic Press. [Chapter 25 is the authoritative description of the Game of Life, including a description of how to embed a universal Turing machine in the Game of Life. For a general audience.]

Boden, M. A. 1999. "Is metabolism necessary?" *British Journal of the Philosophy of Science* 50: 231–48. [Distinguishes three senses of metabolism and examines their implications for strong artificial life. For postsecondary school audiences.]

Dawkins, R. D. 1983. "Adaptationism was always predictive and needed no defense." *Behavioral and Brain Sciences* 6: 360–1. [A defense of adaptationism. For postsecondary school audiences.]

Dennett, D. C. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster. [An extended essay on how natural selection transforms our view of humanity's place in the universe. Accessible to a general audience.]

Gould, S. J. 1989. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: Norton. [A defense of radical contingency in the evolution of life, in the context of detailed examination of fossils in the Burgess shale. Accessible to a general audience.]

—— and Lewontin, R. C. 1979. "The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme." *Proceedings of the Royal Society B* 205: 581–98. [The classic criticism of the use of adaptive explanations in biology. For postsecondary school audiences.]

Harnad, S. 1994. "Levels of functional equivalence in reverse bioengineering." *Artificial Life* 1: 293–301. [Distinguishes synthetic and virtual artificial life, argues that virtual artificial life is impossible, and suggests a Turing test for life to settle whether synthetic artificial life is possible. For postsecondary school audiences.]

Harré, Rom. 1985. *The Philosophies of Science*. Oxford: Oxford University Press. [An introduction to the philosophy of science, for secondary and postsecondary school audiences.]

Kauffman, S. A. 1993. *The Origins of Order: Self-organization and Selection in Evolution*. New York: Oxford University Press. [A technical presentation of Kauffman's views of the place of spontaneous order in living systems. For postgraduate audiences.]

——. 1995. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. New York: Oxford University Press. [Kauffman's views presented to a general audience.]

Kim, J. 1999. "Making sense of emergence." *Philosophical Studies* 95: 3–36. [An investigation of the credibility of emergence with special reference to the philosophy of mind. For postsecondary school audiences.]

Langton, C. G., ed. 1989. *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*. Redwood City: Addison-Wesley. [Proceedings of the first "artificial life" conference identified as such. The editor's introduction (expanded in Boden 1996) is a classic introductory overview of the field. Includes 25 technical papers and a 40-page annotated bibliography of works relevant to artificial life. For a general audience.]

Maynard Smith, J. 1975. *The Theory of Evolution*, 3rd ed. New York: Penguin. [A classic, for a general audience.]

Mayr, E. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge, MA: Harvard University Press. [A history of the ideas in biology, for postsecondary school audiences.]

Pattee, H. H. 1989. "Simulations, realizations, and theories of life." In C. G. Langton, ed., *Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems*. Redwood City: Addison-Wesley, pp. 63–78. [A criticism of strong artificial life. For postsecondary school audiences.]

Ray, T. S. 1992. "An approach to the synthesis of life." In C. Langton, C. Taylor, D. Farmer, and S. Rasmussen, eds., *Artificial Life II*. Redwood City: Addison-Wesley, pp. 371–408. [The classic discussion of one of the best-known artificial-life models. For postsecondary school audiences.]

Schrödinger, E. 1969. *What is Life?* Cambridge: Cambridge University Press. [The classic statement of the metabolic view of life. For a general audience.]

Sober, E. 1992. "Learning from functionalism – prospects for strong artificial life." In C. Langton, C. Taylor, D. Farmer, and S. Rasmussen, eds., *Artificial Life II*. Redwood City: Addison-Wesley, pp. 749–65. [Explores what might be concluded about strong artificial life from recent related work in the philosophy of mind. For postsecondary school audiences.]

Sperry, R. W. 1969. "A modified concept of consciousness." *Psychological Review* 76: 532–6. [A classic psychological presentation of consciousness as an emergent phenomenon. For postsecondary audiences.]

Von Neumann, J. 1966. *Theory of Self-reproducing Automata*. Urbana-Champaign: University of Illinois Press. [Von Neumann's classic work on self-reproducing automata, completed and edited after his death by Arthur Burks. For postgraduate audiences.]

Wiener, N. 1948. *Cybernetics, or Control and Communication in the Animal and the Machine*. New York: Wiley. [The classic work on cybernetics.]

Wolfram, S. 1994. *Cellular Automata and Complexity*. Reading: Addison-Wesley. [A collection of technical papers on cellular automata as complex systems. For postsecondary school audiences.]

Part V

Language and Knowledge

Chapter 16

Information and Content

Jonathan Cohen

Mental states differ from most other entities in the world in having semantic or intentional properties: they have meanings, they are about other things, they have satisfaction- or truth-conditions, they have representational content. Mental states are not the only entities that have intentional properties – so do linguistic expressions, some paintings, and so on; but many follow Grice (1957) in supposing that we could understand the intentional properties of these other entities as derived from the intentional properties of mental states (viz., the mental states of their producers). Of course, accepting this supposition leaves us with a puzzle about how the non-derivative bearers of intentional properties (mental states) could have these properties. In particular, intentional properties seem to some to be especially difficult to reconcile with a robust commitment to ontological naturalism – the view that the natural properties, events, and individuals are the only properties, events, and individuals that exist. Fodor puts this intuition nicely in this oft-quoted passage:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep . . . If aboutness is real, it must be really something else. (Fodor 1987: 97)

Some philosophers have reacted to this clash by giving up one of the two views generating the tension. For example, Churchland (1981) opts for intentional irrealism in order to save ontological naturalism, while McDowell (1994) abandons naturalism (in the sense under discussion) in favor of a kind of intentional dualism as a way of preserving intentional realism. (Terminological caution: unfortunately, McDowell reserves the term 'naturalism' for his own view, and refers to the sort of naturalism under discussion here – which he rejects – as 'bald naturalism'.) However, others hope to find a way of reconciling their naturalistic ontology with intentional realism. In particular, many propose to locate intentionality within the natural order by *naturalizing the intentional* – by offering an account of intentional features in naturalistically respectable terms.

Many philosophers pursuing this project think that an appeal to *information* might satisfy their needs – they believe that this notion is both naturalistically acceptable and adequate to the analysis of the intentional. In this chapter I shall review several attempts to naturalize the intentional in terms of information. In section 1 I'll lay down some conditions of adequacy on informational theories of content that will be useful in evaluating the theories presented later. Next,

in section 2, I'll consider Dretske's influential early formulation of an informational theory. I'll move on in sections 3–4 to discuss some elaborations of the view that incorporate notions of epistemic optimality and teleology. Finally, in section 5 I'll discuss another informational theory, due to Fodor, that turns on his notion of asymmetric dependence. (Please note that there is a short glossary of technical terms at the end of the chapter.)

1 Adequacy Conditions

It will be useful to structure the discussion around a number of adequacy conditions that an acceptable naturalistic theory of content must meet:

Naturalism: To vindicate the naturalism in whose service it is deployed, an informational theory of content must show how intentional properties can be characterized in naturalistic terms. While it is not obvious what terms should count as acceptably naturalistic, proposed naturalistic accounts of intentional properties should not *presuppose* that intentional properties are compatible with naturalism (on pain of begging the very question such theories are intended to answer). Consequently, the necessary and sufficient conditions for having intentional features required by a naturalistic theory of content must not themselves employ intentional idioms, nor should they employ any other notions whose compatibility with naturalism is in doubt. That is to say, the naturalism constraint is intended to force analyses of intentional features to break out of the circle of intentional features, and therefore proposed analyses of these features in terms of others of these features – satisfaction, truth, content, believing that *p*, etc. – will count for present purposes as unacceptably non-naturalistic. On the other hand, the insistence on naturalism is not intended to preclude abstract entities (e.g., laws, properties, numbers) from appearing in the analysis of content.

Grain: Since at least Frege (1892), philosophers of mind and language have made much of the extremely fine grain of individuation for content. In particular, there can be distinct contents that are equivalent in any of a number of senses (e.g., logical, analytic, metaphysical, or nomic equivalence). Many follow Frege in arguing for this conclusion from the observation that it is possible rationally to believe that *p* and disbelieve that *q*, even when *p* and *q* are equivalent in one of these ways; for example, even assuming that it is metaphysically necessary that the Morning Star is identical to the Evening Star, it is possible rationally to believe that the Morning Star is wet and simultaneously disbelieve that the Evening Star is wet (you might be in such a state if you didn't know that the Morning Star is identical to the Evening Star). A theory of content, therefore, must make it possible to distinguish between contents that are logically, analytically, metaphysically, or nomically equivalent.

Misrepresentation: An acceptable theory of intentional content must make it possible that intentional contents can represent the world erroneously – that they can represent the world as being a certain way even when, in fact, the world is not that way (the problem of formulating an account that makes room for misrepresentation is sometimes called "the problem of misrepresentation" or "the problem of error").

2 Dretske and the Flow of Information

One of the earliest systematic attempts to understand intentional content in terms of information occurs in Fred Dretske's seminal *Knowledge and the Flow of Information* (1981). (Dretske's was not, however, the first such proposal; antecedents include Stampe 1975 and 1977, and some suggestive remarks about natural meaning from Peirce 1931 and Grice 1957.) (Dretske's account is also discussed in Chapter 17, KNOWLEDGE.) Dretske wants to understand the intentional content of a signal (e.g., a mental state) in terms of the information that that signal carries under certain circumstances. He understands information in terms of objective conditional probabilities between events (I shall ignore problems concerning the interpretation of these probabilities raised in Loewer 1983 and 1987): he writes

(p. 65) that a signal *r* carries the information that *p* just in case the conditional probability of *p*, given *r* (and *k*, the knowledge of the receiver of *r*), is 1 (but given *k* alone, less than 1).

But we cannot straightforwardly identify information carried with intentional content because a signal will usually carry too much information. For example, an acoustic signal carrying the information that the doorbell is ringing will typically also carry the information that the doorbell's button is being pressed. In contrast, it seems that intentional content is more constrained: I can have a belief that the doorbell is ringing without having the belief that the doorbell's button is being pressed. Therefore, to give an account of intentional content, Dretske needs to rule out the sort of informational nesting found in the example described. To do this, he stipulates that the information that *p* is *nested* in the information that *q* just in case *q* carries the information that *p* (71), and then claims that a signal S has the fact that p as its semantic content iff:

1. *S* carries the information that *p*, and
2. *S* carries no other information, *q*, such that the information that *p* is nested (nomically or analytically) in *q*. (p. 185)

Dretske emphasizes that his notion of information transmission – hence also his notion of intentional content – presupposes a counterfactual-supporting connection between the signal and the information it carries. As a result, a signal correlated with *p* will fail to carry the information that *p* if the correlation is merely accidental or statistical: my thermometer carries information about the temperature of my room and not yours, even if the two rooms are the same temperature, because the state of my thermometer supports counterfactuals about the temperature of my room but not about the temperature of your room (that is to say, it is a true generalization that if the temperature of my room were different, the state of my thermometer would be different; in contrast, it is not generally true that if the temperature of your room were different, the state of my thermometer would be different).

The view of intentional content set out so far is plausibly thought of as meeting the naturalism requirement (putting aside worries about its interpretation of probabilities raised by Loewer – see above), but it faces challenges concerning the desiderata of grain and misrepresentation.

First consider the problem of grain. While Dretske's requirement of counterfactual support allows him to set aside merely correlated events in determining the content of a signal, it will not allow him to choose between properties whose covariation is necessary. (I assume, following Dretske, that our underlying theory of event individuation allows for a distinction between instantiations of properties that necessarily covary.) Suppose that a signal carries the information that *p*, but that it is (nomically, metaphysically, analytically, or logically) necessary that *p* covaries with a distinct property *q*; in any of these cases, it will be nomically necessary that *p* covaries with *q*. (A special case comes from Quine's famous "gavagai" puzzle from Quine 1964.) Quine argued that if a field linguist encountered the term "gavagai" in an unfamiliar language and noted that natives assented to the use of this term when and only when rabbits were present, there would be no fact of the matter which of many incompatible English translations of the term is correct. Live possibilities for the translation, according to Quine, include the following: "rabbit," "undetached rabbit part," "instantaneous temporal stage of a rabbit," "instance of the universal *rabbithood*," and "part of the scattered mereological sum of all rabbits." This provides a special case of the problem under discussion because it is necessary that the properties picked out by these expressions covary. (See the discussion of the "gavagai" puzzle in the context of informational theories of content in Gates 1996.) In this case, the signal will also carry the information that *q*. Presumably the intentional content that *p* can be distinct from the intentional content that *q*; so which is the intentional content of the signal? Dretske is prepared to admit that a signal cannot have one of these contents without having the other (p. 264 n. 2).

This admission has struck many as counterintuitive; however, the problem is even more serious than Dretske's admission would suggest. In fact, Dretske's account has the result that (not both, but) *neither* of the two pieces of information considered can be the content of any signal. In the case described, the information that *p* is

nomically nested in the information that *q*, and the information that *q* is nomically nested in the information that *p*. But on Dretske's account, the intentional content of a signal cannot be a fact that is nomically nested in some other piece of information carried by that signal, so no signal can have either the content *p* or the content *q*. (It is left open that a signal could have the disjunctive content [p ∨ q].) This problem is obviously quite general, and so is a serious objection against Dretske's account.

Second, it seems that the account spelled out so far cannot accommodate the possibility of misrepresentation. This is because, according to that account, a signal *S* with the intentional content *p* must carry the information that *p*, and this requires that the conditional probability of *p* given *S* is 1 – i.e., *p* must be true. (Cf. the discussion of the so-called "disjunction problem" in Fodor 1990d: if both *p* and *q* can cause *S*s, how can a theory of content make it the case that *q*-caused *S*s have the erroneous content *p* rather than the always veridical disjunctive content [p ∨ q]?) Thus, on the theory we have considered so far, intentional contents cannot misrepresent the world.

Dretske is aware of the problem of misrepresentation, and attempts to answer it by proposing that only some of the tokenings of a signal carry the information that determines that signal's content. In particular, he proposes that there is a learning period for a signal, during which that signal carries the information that *p*, and that the signal's intentional content is given only in terms of the information it carries in its learning period. After the learning period, when the intentional content of the signal has already been fixed as *p*, tokenings of that signal can fail to carry the information that *p*, and so can be erroneous:

> In the learning situation special care is taken to see that incoming signals have an intensity, a strength, sufficient unto delivering the required piece of information *to* the learning subject . . . But once we have meaning, once the subject has articulated a structure that is selectively sensitive to [the information that *p*] . . . , instances of this structure, tokens of this type, can be triggered by signals that lack the appropriate piece of information. (pp. 194–5)

Dretske's answer to the problem of misrepresentation raises a number of problems of its own. First, it has seemed to many implausible that there is anything like a principled distinction between the learning period and the nonlearning period for most signals. A second concern is that, even if there is a learning period for signals, this period must be characterized nonintentionally if the naturalism constraint is to be respected. This requirement precludes understanding the learning period for a signal simply as the period leading up to that signal's having the intentional content *p*, and it is not obvious that there is an alternative naturalistically acceptable understanding in the offing. Third, relying on the learning period to explain misrepresentation leaves Dretske without an account of how unlearned (innate) signals could misrepresent. Fourth, as Loewer 1997 notes, the account in terms of a learning period is implausible for many signals; for instance, it seems possible that a child could learn that the linguistic symbol "aardvark" has aardvarks (not pictures of aardvarks) as its content, even if all tokens of "aardvark" in the learning period carry information about pictures of aardvarks rather than aardvarks.

3 Epistemic Optimality

Many are convinced by consideration of these difficulties that content can't be reconstructed in terms of information alone. However, many believe these problems can be solved by a hybrid theory that appeals to both an informational factor and some other (not strictly informational) factor. The thought is that, so long as both the informational and the non-informational factors can be given naturalistic explications, they can work together to provide a naturalistic account of content that escapes the vulnerabilities of more strict informational theories.

One family of theories of this sort appeals to a notion of epistemic optimality to take up the slack left by information. On these views, a signal *S* has the content *p* iff there are epistemically optimal conditions C_p for *p* such that if C_p obtained, then *S* would nomologically covary with *p*.

Proponents of such accounts, including Stampe (1975, 1977), Dretske (1983), and Fodor (1990b), hope that their appeal to epistemic optimality might resolve the problem of misrepresentation; this thought is motivated by the (reasonable) suggestion that misrepresentation occurs when cognitive systems attempt to represent the world while operating in epistemically suboptimal conditions. Thus, for example, tokens of the linguistic symbol "aardvark" that are caused by armadillos seen on dark nights have as their content the property *aardvark* – rather than *armadillo on a dark night* or *aardvark or armadillo on a dark night* – because the symbol "aardvark" would covary only with instances of *aardvark* in epistemically optimal conditions (epistemically optimal conditions for *aardvark* presumably involve the lighting being up, the subject's attentively looking in the right direction, and so on).

It has also been suggested that appeals to optimality could solve the problem of grain. For even if p and q covary, we could say that a symbol has p rather than q as its content if S would nomologically covary with p but not q in epistemically optimal conditions for p. However, it is unclear that this appeal to epistemic optimality resolves the problem of grain. For one thing, this solution will fail if the optimal conditions for p (C_p) are identical to the optimal conditions for q (C_q); for in this case, S would, once again, nomologically covary with both p and q in C_p (and C_q, of course), and so would not determinately have the content p. A variant of this worry arises when "p iff q" is necessary. For, here again, S would covary with both p and q in all possible conditions, *a fortiori* in C_p (assuming condition C_p is possible; if not, then it would not be true that S nomologically covaries with p in C_p, so S could not have the content that p).

Moreover, the naturalistic credentials of epistemic optimality theories are questionable as well. This can be seen in two ways. First, it is plausible that the epistemic optimality conditions for a content p cannot be stated without adverting to the content p itself, since what is epistemically optimal seems to depend on what content we're hoping to reconstruct. For example, the optimality conditions for the belief that there's an aardvark in the room preclude looking through a microscope, but the optimality conditions for the belief that there's a paramecium in the room require looking through a microscope. But, of course, epistemic optimality versions of informational theories explain the content of signals in terms of the optimality conditions for that content. Consequently, the understanding of a signal's having the content p provided by an epistemic optimality theory must be stated in terms of the content p, contrary to the naturalism requirement. Second, insofar as belief fixation is widely thought to be a holistic enterprise, it is equally plausible that the epistemic optimality conditions for a content p cannot be stated without adverting to contents other than p. For example, my tokenings of the linguistic symbol "there are aardvarks in the room" won't nomologically covary with the presence of aardvarks in the room if I believe that aardvarks are not macroscopically observable. But if so, then an epistemic optimality theory's unpacking of what it is for a signal to have the content that p must advert to states that must be characterized by their contents. And once again, this seems a clear violation of the naturalism constraint.

4 Teleology

An alternative elaboration of the informational approach to content, appealing to considerations of teleology, rather than epistemic optimality, is advocated by several philosophers (see Fodor 1984 and 1990b, Millikan 1984, 1986, and 1989, Dretske 1988, Papineau 1993). Here, too, the hope is that the theory's non-informational factor – its appeal to teleology – will resolve the problems of misrepresentation and grain that plague stricter informational accounts. Roughly put, the thought is that misrepresentation occurs when a signal S covaries with the information that q even though S's teleological function is to carry the information that p ($p \neq q$). Similarly for the problem of grain: we could say that a symbol means p rather than q (even if p is satisfied when and only when q is satisfied) if the teleological function of the symbol is to carry the information that p. (These formulations assume – controversially – that teleological functions are

assigned to individual signals, rather than the whole cognitive systems sustaining the signal-world covariations taken as basic by informational theories. Unfortunately, I cannot examine the alternative construal here for reasons of space.)

Proponents of teleological accounts need to explain the notion of teleological function naturalistically. Of course, they cannot appeal to the well-understood example of the teleological function of artifact symbols in this context (e.g., the symbols on the face of a pressure gauge have the teleological function of representing pressure), since these instances of teleological function are presumably constituted in terms of content: the gauge has the function of measuring pressure because that's what its makers intended it to do (for similar reasons, appeals to God's intentions to fix teleological functions are off limits to would-be naturalists as well). Instead, these theorists typically propose understanding teleological function in terms of natural selection (see Wright 1973, Millikan 1984, Neander 1991). On this view, a signal *S* of type *S** has the teleological function of carrying the information that *p* (in an organism *o*) just in case earlier tokens of type *S** were selected (in *o*'s species) by natural selection because they carried the information that *p* – that is, *S* has the function of carrying the information that *p* in *o* just in case the carrying of the information that *p* by earlier tokens of type *S** increased the fitness of *o*'s ancestors.

To be sure, questions remain about the naturalistic *bona fides* of the account just sketched – for example, some object that this formulation buys its naturalism at the price of an implausibly robust conception of natural selection. Moreover, the success of this account depends on the possibility of a naturalistic explanation of how tokens are assigned to signal types. However, the most important objections against teleological accounts allege that they cannot accommodate the desiderata of grain and misrepresentation.

The problem of grain for teleological theories is almost invariably presented in connection with the frog's capacity to snap at flies. Consider an internal state *S* in the frog that covaries with the presence of flies and mediates his snapping behavior. It may be that the frog's environment *E* is such that all the local small moving black objects are flies, all the items of frog-food are flies, all the flies-or-bee-bees are flies, and so on. Then state *S* covaries not only with the presence of flies, but also with the presence of small moving black objects in *E*, frog-food in *E*, flies-or-bee-bees in *E*, and so on. Indeed, state *S* carries information about all of these; so which (if any) is the content of *S*? According to teleological theories, the information that *p* is the content of *S* iff the carrying of the information that *p* by other tokens of the same type increased the fitness of the frog's ancestors. But a token's carrying information about any of the candidates considered above would have an equal effect on the fitness of ancestral frogs: snapping at flies, small moving black objects in *E*, frog-food in *E*, and flies-or-bee-bees in *E* would all get exactly the same things into the frog's belly so long as all the small moving black objects (/frog-food/ flies-or-bee-bees) in *E* are flies. As Fodor puts the point,

> it's equally OK with Darwin which way you describe the intentional objects of y snaps, so long as it's reliable (say, nomologically necessary; anyhow, counterfactual supporting) that all the local flies-or-bee-bees are flies. The point is, of course, that if all the local flies-or-bee-bees are flies, then it is reliable that the frog that snaps at one does neither better nor worse selection-wise than the frog that snaps at the other. (Fodor 1990d: 73)

This problem is even more pressing when the covariation between candidate contents is not merely nomologically necessary, but metaphysically necessary: it is extremely difficult to see how natural selection could favor a signal with the content *aardvark there* over one with the content *undetached aardvark part there*.

The problem of misrepresentation for teleological theories is a consequence of the problem of grain. As we have seen, the teleologist's appeal to natural selection leaves it open that *S* could mean *fly-or-bee-bee* rather than *y*. But if so, then nothing could make it the case that bee-bee-caused *S*-tokens are erroneous rather than veridical. More generally, if this is right, it is difficult to see what could make it the case that any content-assignment by a teleological theory is erroneous.

5 Asymmetric Dependence

Fodor has proposed another informational account as a way of giving a semantics for expressions in a language of thought (henceforth, Mentalese). (Fodor has presented significantly different versions of the theory over the years; see Fodor 1987, 1990, 1990, 1994. For reasons of space, I confine myself here to the formulation in Fodor 1990e, his most complete statement of the view.) The intuition behind this view is that, while a Mentalese symbol can carry information about a number of distinct properties, only one of them is its content – namely, its content is that single property on which its carrying information about all the other candidate properties depends. For example, the intuition runs, while my mental state can carry information about both *aardvark* and *armadillo on a dark night*, its content is *aardvark* because it would not carry information about *armadillo on a dark night* (or anything else) unless it carried information about aardvark. This intuition is fleshed out in Fodor 1990e in terms of a relation of asymmetric dependence between laws, which is itself specified in terms of a pair of subjunctive conditionals: a law L_1 is said to depend asymmetrically on a law L_2 just in case (i) if L_2 did not hold, then L_1 would not hold either, and (ii) if L_1 did not hold, then L_2 would still hold. (That said, Fodor sometimes suggests that the asymmetric dependence relations are metaphysically basic and not in need of cashing in terms of subjunctive conditionals; see Fodor 1990e: 93, 95.)

Using this apparatus, Fodor writes (p. 121) that a Mentalese expression *S* has property *p* as its content if:

1. "*p*s cause *S*s" is a law.
2. Some *S*s are actually caused by *p*s.
3. For all $q \neq p$, if *q*s qua *q*s actually cause *S*s, then *q*s causing *S*s is asymmetrically dependent on *p*s causing *S*s.

This formulation deserves supplementation. (In what follows I shall use capital letters to indicate Mentalese expressions: AARDVARK is a Mentalese expression that, we hope, has the property *aardvark* as its content.) First, Fodor understands the laws appealed to in this account as *ceteris paribus* generalizations – generalizations that are true (and counterfactual-supporting) when other things are held equal. As a result, the proposal is not vulnerable to potential counterexamples involving the failure of aardvarks to cause AARDVARKs in dark rooms, for inattentive subjects, and so on: these would be brushed aside as cases where the relevant *paria* aren't *cetera*, and therefore as falling outside the intended scope of the laws. Second, Fodor insists that the subjunctive conditionals in terms of which asymmetric dependence is characterized should be read synchronically rather than diachronically (p. 134 n. 18). That is, it might be that aardvark-pictures played a diachronically ineliminable role in the ontogeny of my thought (say, because I was taught to token AARDVARK exclusively by a process involving exposure to aardvark-pictures); in this case, aardvarks wouldn't cause AARDVARKs in me were it not that aardvark-pictures had first caused AARDVARKs in me. Still, according to Fodor, the aardvark-picture to AARDVARK nomic link is asymmetrically dependent on the aardvark to AARDVARK nomic link because, he thinks, the *synchronic dependence* goes in the opposite direction from the diachronic dependence: once the capacity to token AARDVARK is in place, aardvark-pictures wouldn't cause AARDVARKs unless aardvarks caused AARDVARKs.

Asymmetric dependence is designed to resolve the problems of grain and misrepresentation. Take the problem of grain first: why is it that AARDVARK means *aardvark* rather than *nocturnal termite-consuming African burrowing mammal*? (For the purposes of illustration, assume it is nomically necessary that these properties covary.) Answer: even if every AARDVARK I token is caused by something that is an instance of both properties, instances of *nocturnal termite-consuming African burrowing* mammal would not cause AARDVARKs unless instances of *aardvark* caused AARDVARKs, while the reverse dependency does not hold. Take the problem of misrepresentation next: given that some AARDVARK tokens are caused by armadillos on dark nights rather than aardvarks, why do those tokens have the erroneous content *aardvark* rather than the veridical content *armadillo* on a dark night (or the veridical

disjunctive content *aardvark or armadillo on a dark night*)? Once again, asymmetric dependence supplies an answer: armadillos on dark nights cause AARDVARKs only because the former are mistakenly identified as aardvarks, so the nomic link between *armadillo on dark night* and AARDVARK is asymmetrically dependent on the nomic link between aardvark and AARDVARK (similarly, the link between the disjunctive property and AARDVARKs is asymmetrically dependent on the link between *aardvark* and AARDVARKs).

Several objections have been leveled against Fodor's theory. Many of these take the form of counterexamples designed to show that Fodor's conditions do not assign the correct content to Mentalese expressions (Cummins 1989: ch. 5, Godfrey-Smith 1989, Baker 1991, Boghossian 1991, Manfredi & Summerfield 1992, Adams & Aizawa 1994). Fodor has responded to many of these criticisms (see Fodor 1990e, the replies of Loewer & Rey 1991, and Fodor 1994). Unfortunately, these controversies are often extremely difficult to assess because it is unclear how we should understand the relevant subjunctive conditionals. (The difficulty in assessing alleged counterexamples to Fodor's theory is even more severe if asymmetric dependence is taken as metaphysically basic – see above.)

A further complication is that Fodor intends his theory to provide only sufficient, and not necessary, conditions for a state's having content. As Fodor construes it, his task is only to show how Mentalese expressions *could* have content – not to show how Mentalese expressions *do* have content. By reducing his aspirations in this way, Fodor forestalls objections about whether his conditions are necessary for a state's having content, and in particular about whether the subjunctive conditionals he invokes are true:

> It's enough if I can make good the claim that "*X*" *would mean* such and such if so and so *were to be* the case. It's not also incumbent upon me to argue that since "*X*" *does* mean such and such, so and so *is* the case. (Fodor 1990e: 96)

Whether or not the points considered so far are damaging, the asymmetric dependence theory faces other important obstacles.

First, it is unclear that Fodor's account is acceptably naturalistic. As noted, Fodor understands the laws on which his account rests as true, counterfactual-supporting, *ceteris paribus generalizations.* However, it seems that the *ceteris paribus* conditions governing the aardvark–AARDVARK link cannot avoid adverting to content for two reasons (this mirrors an objection discussed in section 3). For one thing, just which *cetera* must remain *paria* to sustain the token–world links the theory requires depends on what that symbol's content is: instances of *aardvark* can't cause AARDVARKs in me if I am looking through a microscope, but instances of *paramecium* can't cause PARAMECIUMs in me unless I am looking through a microscope. For another, just which *cetera* must remain *paria* to sustain the token–world links the theory requires depends on what other contents the subject believes: I won't reliably token AARDVARK in the presence of aardvarks if I believe that aardvarks are not macroscopically observable. For both these reasons, it seems that the *ceteris paribus* clauses appearing in the laws required by the theory ultimately must be cashed in terms of content, and consequently that the theory cannot be stated without adverting to content. If so, then the view does not abide by the naturalism constraint.

Finally, while the asymmetric dependence account is arguably successful in dispatching many instances of the grain problem, it remains helpless to mark distinctions in content between properties whose covariation is metaphysically necessary. Because it is metaphysically necessary that instances of *aardvark* covary with instances of *undetached aardvark part*, there is a counterfactual-supporting generalization linking instances of *aardvark* to AARDVARK tokens iff there is a counterfactual-supporting generalization linking instances of *undetached aardvark part* to AARDVARK tokens. Moreover, since the covariation between the two properties is metaphysically necessary, this will be so in every metaphysically possible world, so there can be no asymmetric dependence between the two laws. (Presumably Fodor would not respond by appealing to asymmetric dependencies in metaphysically impossible worlds, since those worlds will include many where the dependency goes in the opposite

direction from that required by the theory.) Consequently, the theory predicts that AARDVARK (i) has the content *aardvark* if and only if it has the content *undetached aardvark part*, or (ii) has the disjunctive content *aardvark* or undetached aardvark part. (In Fodor 1994: ch. 3 he recognizes this problem, and suggests that a theory of content can settle on a determinate and nondisjunctive content for a Mentalese expression by appeal to the inferential relations of thoughts containing that expression. This solution is criticized in Gates 1996 and Ray 1997.)

6 Conclusion

Many contemporary philosophers believe that informational theories are the most promising proposals for reconciling naturalism with intentional realism. However, it remains to be shown that there is an informational theory of content that satisfies the constraints of section 1 above. Of course, this does not mean that no informational theory can succeed. It does mean that, so far, appeals to information have not resolved the problem of naturalizing content.

Glossary of Key Technical Terms

analytic necessity: a sentence is said to be analytically necessary if it is necessary in virtue of its meaning. For example, many believe that the sentence "bachelors are unmarried" is analytically necessary. Compare with metaphysical and nomic necessity.

carry (*information*): a signal *r* carries the information that *p* just in case the conditional probability of *p*, given *r* (and *k*, the knowledge of the receiver of *r*), is 1 (but given *k* alone, less than 1) (Dretske 1981: 65).

ceteris paribus: Latin for "other things equal." *Ceteris paribus* generalizations are nonstrict generalizations – generalizations that hold when other things are equal.

covariation (*properties*): properties *P* and *Q* covary just in case *P* is instantiated if and only if *Q* is instantiated.

counterfactual support: a generalization is said to be counterfactual-supporting if it is not only true of its instances, but also would be true of relevant noninstances. For example, the generalization "metals expand on heating" is counterfactual-supporting because it not only says something true about what happens to heated metals, but also says something true about what would happen to unheated metals were they heated (contrary to actual fact). In contrast, "everything in Nelson Goodman's pocket on VE Day was silver" is not counterfactual-supporting because, even if all the things in Nelson Goodman's pocket on VE Day were in fact silver, there are many nonsilver objects that might have been in Nelson Goodman's pocket on VE Day. It is widely thought to be a requirement on nomic generalizations (as opposed to mere accidentally true generalizations) that they support counterfactuals.

holism of belief fixation: the claim that it is impossible to fix, or come to, a given belief without holding in place a number of other beliefs at the same time. According to this view, an experimental datum confirms (i.e. verifies, gives us some reason to believe) a given statement only in conjunction with one's other theoretical commitments, background assumptions about the experiment, and assumptions about the logical and mathematical apparatus connecting the datum to these other beliefs.

intentional: about something. Things that are about other things (e.g., mental states, words) are said to have intentional properties. Not to be confused with a different, ordinary usage of "intentional" to mean on purpose.

Mentalese: language of thought. According to Fodor (1975), thinking should be understood as manipulation of Mentalese symbols with both syntactic and semantic (intentional) properties.

metaphysical necessity: a proposition (or a sentence expressing a proposition) is said to be metaphysically necessary just in case it is necessary by virtue of metaphysical truths. For example, if the correct metaphysics of the constitution of water says that water is H_2O, then it is metaphysically necessary that non-H_2O stuff – even if clear, potable, odorless, tasteless, etc. – is not water. Compare with analytic and nomic necessity.

naturalism: the view that the natural properties, events, and individuals are the only properties, events, and individuals that exist.

nesting: the information that *p* is nested in the information that *q* just in case *q* carries the information that *p* (Dretske 1981: 71). Specifically, *p* is nomically nested in *q* just in case it is nomically necessary (necessary by virtue of natural laws) that *p* is nested in *q*. Similarly, *p* is analytically nested in *q* just in case the sentence "*p* is nested in *q*" is analytically necessary (necessary in virtue of its meaning).

nomic: law-governed.

nomic necessity: a proposition (or a sentence expressing a proposition) is said to be nomically necessary just in case it is necessary by virtue of natural laws. For example, the proposition that metals expand when heated is nomically necessary. Compare with analytic and metaphysical necessity.

subjunctive conditional: a conditional is a hypothetical statement of the form "if *p* then *q*"; the component *p* is called the antecedent of the conditional, while component *q* is called the consequent. Conditionals whose antecedents are in the grammatical subjunctive mood neither presuppose that their antecedents are true nor that they are false; these are called subjunctive conditionals. For example, "if I were to eat a bagel, then I would be full" is a subjunctive conditional: its antecedent is in the subjunctive mood, and it presupposes neither that I do eat a bagel, nor that I do not.

References

Adams, F. and Aizawa, K. 1994. "Fodorian semantics." In S. P. Stich and T. A. Warfield, eds., *Mental Representation: A Reader*. Oxford: Blackwell, pp. 223–42. [Adams and Aizawa argue that there is no tenable reading of Fodor's asymmetric dependence account by proposing counterexamples (cases where the theory assigns the wrong content, or assigns content where it should not). They also argue that Fodor gets the order of explanation wrong – that asymmetric dependence is a consequence of intentional relations, not the ground of those relations.]

Baker, L. R. 1991. "Has content been naturalized?" In Loewer & Rey 1991: 17–32. [Baker argues that Fodor vacillates between different (and inconsistent) versions of his theory as he attempts to handle different problem cases, but that no version is adequate as a general solution.]

Boghossian, P. A. 1991. "Naturalizing content." In Loewer & Rey 1991: 65–86. [Boghossian argues that Fodor's asymmetric dependence view collapses onto an optimal conditions theory, and that it is therefore susceptible to the kinds of objections that Fodor himself raises against the latter view.]

Churchland, P. M. 1981. "Eliminative materialism and the propositional attitudes." *The Journal of Philosophy* 78: 67–90. [Churchland argues that "folk psychology" – the commonsense predictive/explanatory understanding of the mental that centrally involves the attribution of intentional states – is a radically and systematically erroneous theory in need of replacement by a scientific psychology pitched at a neural level.]

Cummins, R. 1989. *Meaning and Mental Representation*. Cambridge, MA: MIT Press. [Cummins discusses a number of (informational and non-informational) approaches to content in this slim book, and presents his own "interpretational" semantics.]

Dretske, F. I. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press. [This is Dretske's first major defense of an informational theory of content (see section 2 above). This book was instrumental in bringing informational approaches to the attention of mainstream philosophers of mind.]

——. 1983. "The epistemology of belief." *Synthese* 55: 3–19. [Dretske argues that if a system cannot process information to the effect that something is *F*, then it cannot represent something as *F*, *a fortiori* cannot believe that something is *F*. He also endorses an epistemic optimality account.]

——. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press. [Dretske aims to explain how reasons, citing psychological states, can figure in the explanation of actions, given the (in principle) availability of complete explanations in terms of physical causes. As part of this project, Dretske defends a historical/teleological information-theoretic account of the intentional.]

Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press. [Fodor

argues that our best psychological theories presuppose the existence of a language of thought (Mentalese) – a language-like system of symbols with syntactic and semantic properties.]

——. 1984. "Semantics, Wisconsin style." *Synthese*, 59: 231–50, Repr. in Fodor 1990c: 31–49. [Fodor criticizes views associated with Dretske and Stampe for failing to solve the disjunction problem.]

——. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press. This book contains Fodor's first published statement of his asymmetrical dependence theory.]

——. 1990a. "Information and representation." In P. Hanson, ed., *Information, Language, and Cognition*. Vancouver: University of British Columbia Press, pp. 175–90. [Fodor marks a distinction between "labeling" uses (e.g., "that is an aardvark") and "representational" uses (e.g., "an aardvark is a beautiful animal") of the same symbol: representational uses are often not intended to indicate that the cause of the symbol has the property expressed by the symbol. He insists that the asymmetric dependence account is intended only to apply to labeling uses of symbols.]

——. 1990b. "Psychosemantics or: Where do truth conditions come from?" In W. G. Lycan, ed., *Mind and Cognition: A Reader*. Malden, MA: Blackwell, pp. 312–38. [Widely circulated as a manuscript circa 1980. Fodor defends an optimality/teleology informational account. He repudiates this view in Fodor 1984, 1990d, and 1987.]

——. 1990c. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press. [This collection contains, among a number of previously published essays, two previously unpublished papers on content that comprise the most worked-out statement of Fodor's asymmetric dependence view.]

——. 1990d. "A theory of content, I: The problem." In Fodor 1990c: 51–87. [Here Fodor discusses the problems that an informational theory of content must solve, and argues that going versions of such theories are unsuccessful at solving them.]

——. 1990e. "A theory of content, II: The theory." In Fodor 1990c: 89–136. [This is Fodor's most complete presentation of his asymmetric dependence theory. Fodor motivates the view, explains and discusses his approach in some detail, and responds to a large number of potential objections.]

——. 1994. *The Elm and the Expert: Mentalese and its Semantics*. Cambridge, MA: MIT Press. [Fodor engages in a number of projects in this book, derived from his Nicod Lectures of 1994. Among these, he offers an answer to Quinean inscrutability problems in chapter 3 and some substantive revisions and extensions of his asymmetric dependence theory in two appendices.]

Frege, G. 1892. "On sense and meaning." In B. McGuinness, ed., *Gottlob Frege: Collected Papers on Mathematics, Logic, and Philosophy*. Oxford: Blackwell, pp. 157–77. [In the opening pages of this paper, Frege notes that co-referential expressions are not intersubstitutable *salva veritate* in propositional attitude contexts. This point is at the heart of the grain constraint discussed in section 1 above.]

Gates, G. 1996. "The price of information." *Synthese* 107(3): 325–47. [Gates argues that informational theories of content presuppose a solution to Quinean inscrutability, and that no such solution is available.]

Godfrey-Smith, P. 1989. "Misinformation." *Canadian Journal of Philosophy* 19: 533–50. [Godfrey-Smith claims that Fodor fails to solve the disjunction problem – that he has no way to ensure that the content of wild tokens will depend asymmetrically on that of veridical tokens, as required by the asymmetric dependence theory.]

Grice, H. P. 1957. "Meaning." *The Philosophical Review* 66: 337–88. [Grice distinguishes between "natural meaning" – the kind of meaning that we are speaking of when we say something like, "Those spots mean measles" – and "non-natural meaning" – the kind of meaning we speak of when we say "Those three rings on the bell mean that he made it safely to shore." He proposes that the non-natural meaning of symbols can be understood in terms of the intentions of the producers of those symbols. Informational theories of content are proposed to explain (as a kind of natural meaning) the content of the intentions constituting non-natural meanings.]

Loewer, B. 1983. "Information and belief." *Behavioral and Brain Sciences* 6: 75–6. [Among other complaints, Loewer's brief comment on Dretske objects that the probabilities to which Dretske appeals cannot be understood in any of the standard ways.]

——. 1987. "From information to intentionality." *Synthese* 70: 287–317. [Loewer argues that extant informational theories are unsuccessful in providing naturalistically acceptable understandings of content.]

——. 1997. "A guide to naturalizing semantics." In B. Hale and C. Wright, eds., *A Companion to the Philosophy of Language*. Oxford: Blackwell, pp. 108–26. [Loewer extends the arguments of Loewer 1987 to the effect that extant informational theories of content cannot meet reasonable desiderata.]

—— and Rey, G., eds. 1991. *Meaning in Mind: Fodor and His Critics*. Oxford: Blackwell. [This book contains a short summary of Fodor's work by the editors, 14 critical essays on various aspects of his *oeuvre* (including several on his asymmetric dependence theory), Fodor's replies to all of these, and an extensive bibliography of Fodor's writings.]

Manfredi, P. and Summerfield, D. 1992. "Robustness without asymmetry: A flaw in Fodor's theory of content." *Philosophical Studies* 66: 261–83. [Manfredi and Summerfield offer another purported counterexample to Fodor's account. They conclude that asymmetric dependence cannot be a general solution to robustness.]

McDowell, J. 1994. *Mind and World*. Cambridge, MA: Harvard University Press. [This book is of interest in the present context principally because it rejects the naturalism constraint I set out in section 1. McDowell believes that the normative dimension of content cannot be captured naturalistically, and therefore is unsympathetic to the goal of naturalizing content (and other forms of "bald naturalism").]

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA: MIT Press. [Millikan presents a general naturalistic theory of teleological function (one that is intended to apply to body organs, behaviors, and mental representations, *inter alia*), and defends a teleological account of content.]

——. 1986. "Thought without laws: cognitive science with content." *The Philosophical Review* 95: 47–80. [Millikan holds that references to the content of psychological states are ineliminable in both folk psychology and cognitive science. However, she thinks, psychology is naturalistic because content can be unpacked by appeal to the evolutionary history of systems and states.]

——. 1989. "Biosemantics." *The Journal of Philosophy* 86: 281–97. Repr. in Stich & Warfield 1994. [Millikan clarifies, elaborates, and defends the teleological theory of content presented in Millikan 1984.]

Neander, K. 1991. "Functions as selected effects: the conceptual analyst's defense." *Philosophy of Science* 58: 168–84. [Neander defends an etiological theory of biological function – a theory on which the function of a trait is the effect for which it was selected by natural selection. This view gives one way of understanding the notion of function that figures centrally in teleological informational theories of content.]

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell. [This book is a very general defense of naturalism in metaphysics and epistemology. Part II spells out a teleological construal of intentional properties that is intended to respect Papineau's naturalist aims.]

Peirce, C. S. 1931. *Collected Papers*. Cambridge, MA: Harvard University Press. [Volume 2 of this work contains a very early formulation of an informational theory of content.]

Quine, W. V. O. 1964. *Word and Object*. Cambridge, MA: MIT Press. [This is the source of the "gavagai" problem that, according to Gates 1996 plagues contemporary informational theories of content.]

Ray, G. 1997. "Fodor and the inscrutability problem." *Mind and Language* 12(3/4): 475–89. [Ray attempts to undermine Fodor's answer to the inscrutability problem in Fodor 1994. He gives several examples to which, he claims, Fodor's answer is inapplicable, and concludes that Fodor's solution cannot secure the scrutability of reference in the general case.]

Stampe, D. 1975. "Show and tell." In B. Freed, A. Marras, and P. Maynard, eds., *Forms of Representation: Proceedings of the 1972 Philosophy Colloquium of the University of Western Ontario*. Amsterdam: North-Holland. [This is one of the earliest formulations of an informational theory of content, aimed primarily at the understanding of linguistic content.]

——. 1977. "Toward a causal theory of linguistic representation." In P. French, T. Uehling, and H. Wettstein eds., *Midwest Studies in Philosophy*, vol. 2. Minneapolis: University of Minnesota Press. [Here Stampe develops and defends the view of Stampe 1975. This paper is widely cited

as a primary source of contemporary interest in informational approaches to content.]

Stich, S. P. and Warfield, T. A., eds. 1994. *Mental Representation: A Reader*. Oxford: Blackwell. [A fine collection of previously published papers on mental content. The papers cover a wide range of (informational and non-informational) approaches.]

Wright, L. 1973. "Functions." *The Philosophical Review* 82: 139–68. [Wright proposes that the analysis of biological functions must make reference to the causal background of the bearers of such functions. He argues that this sort of "etiological" account avoids difficulties that plague competing analyses and secures several important desiderata for the notion of function. Proponents of teleological accounts of content have often relied on accounts of function in this tradition.]

Chapter 17

Knowledge

Fred Adams

Introduction

In business and finance, "managing information" and "managing knowledge" practically are synonymous. Surfing internet search engines provides confirmation. What is it about knowledge and information that makes the two suited to one another? At the level of common sense, if one possesses the information that *p*, one is in a position to know that *p*, all else being equal. And if one is uninformed, one is not in a position to know that *p*. Similarly, if one knows that *p*, one has more than a justified true belief (Gettier 1963) – justified beliefs can be true accidentally. I see Ken and believe I do (someone I know well). But my belief that it is Ken is accidentally true and not knowledge because, although I see Ken, I cannot distinguish Ken from his identical twin Ben, who is standing right behind Ken. Information that it is Ken may secure the needed connection between belief and truth to take me beyond a merely true (even justified) belief that I see Ken. Whether this is true beyond the level of common sense depends upon a deeper understanding of knowledge and information.

Let's begin with knowledge. It is uncontroversial that knowledge requires truth and belief. One cannot know that Bush got more votes in Florida than Gore in the 2000 American Presidential election, if it is not true, though one can certainly believe it. And one cannot know that the US Supreme Court stopped the statewide Florida hand recount, if one has no beliefs about it whatsoever, though it can be true. So truth and belief, though independent, are both required for knowledge. Knowledge requires that these two things, otherwise independent, come together.

Nozick (1981) claims that knowledge tracks the truth, such that the beliefs of a knower would co-vary with the truth – when one applies the right belief-forming methods. Figuratively, a knower's cognitive light of belief would illuminate, were *p* true, and not illuminate, were *p* false.

Attention turns to methods. What sorts of methods would provide the mechanisms to implement the sort of tracking of truth required by knowledge on a theory like Nozick's? For knowledge of empirical propositions, one would expect to need empirical methods: perceptual observation, instrumentation, testimony, or perhaps an entire empirical theory. Non-empirical truths of logic or mathematics may require mathematical intuition, reason (a form of cognitive computation), formal systems, and techniques of proof.

Justification is also a notion that is traditionally associated with knowledge, for it is often thought to be a necessary ingredient in getting beliefs to track the truth in ways capable of generating knowledge. There is a wide variety of theories of justification, and each emphasizes different aspects

or factors relevant to knowing. Externalists stress factors that are possibly outside the mind or awareness of the believer. Internalists stress factors of which a believer must be aware. Foundationalists attend to the origin, grounds, and structure of support for beliefs that are not inferred from other beliefs. Coherentists stress the interrelation of beliefs and their relations of mutual confirmation. And reliabilists attend to the likelihood that a set of cognitive mechanisms and their conditions of operation tend to produce true beliefs more often than false beliefs. Despite their differences, all these theories agree that a belief's being justified contributes to minimizing that a belief is merely true, but not known. The exact relation between a belief's being justified and its being true continues to be difficult to specify with precision, partly because there can be justified false beliefs. Still, it is undeniable that epistemic justification is prized because it either increases the likelihood that one non-accidentally has a true belief, or, having that, it satisfies an additional cognitive requirement of knowing (coherence, evidence, or rationality).

Some have argued that if justification is something other than information (Adams 1986) or other than what, when added to true belief, guarantees that the belief is non-accidentally true (Lewis 1996), then justification is neither necessary nor sufficient for knowledge. Still, a theory of knowledge that incorporates information might tack one or other account of justified belief onto its conditions of knowing. Alternatively, one might attempt to substitute information for justified belief into one's account of knowledge. In what follows, we will consider an information-theoretic attempt to do an "end run" around giving a theory of justified belief.

Now let's turn to information. To be of value to a would-be knower, information must be an objective, mind-independent commodity. In principle, it should be possible for someone to be the first person to learn that *p*. If *S* were the first person brought to know that *p* by the information that *p*, then the information that *p* would appear to have objective properties. The following examples suggest that this is so. Waves of radiation traveling through space may contain information about the Big Bang before anyone detects it. Fingerprints on the gun may contain information about who pulled the trigger before one lifts the prints. Thus, information appears to be mind-independent (and, thereby, language-independent too).

Information must also be capable of having a very special relationship to the truth. Since one cannot know what is false, if information is going to bring one to know that *p*, then information must also be tied to the state of affairs that makes *p* true. Otherwise, it is hard to see the value of information over belief and truth itself. On at least some accounts, information has this connection to truth (Dretske 1981, Floridi forthcoming *b*). One can be misinformed. One can be informed that *q*, when one needs to know that *p*, but one cannot be misinformed that *p*. For something can only carry the information that *p*, if *p*. Indeed, if we think of information as being contained or carried in one event (set of events) and as being about another event (set of events), then the transmission of information is the product of a correlation and dependency between the two events (sets). To see this in more detail, let's consider Dretske's (1981) attempt to explicate an account of information that may be useful in understanding knowledge.

We will first look at Dretske's account and see how he uses information to explicate knowledge. We will also look at some interesting objections to his account. This will give us a good idea of the usefulness of information in understanding what knowledge is. We will then consider some interesting open questions about knowledge and information. And we will close with a survey of some current philosophical debates about knowledge.

Dretske's Adaptation of Information Theory to Knowledge

To adapt information theory to a format friendly to a theory of knowledge, several matters need to be resolved. For example, to know that Bush was elected president involves information being generated by the event of his election. It also involves transmission of that information to a prospective knower *S*. *S* must detect physical events that carry that transmitted information,

and those events must cause or sustain *S*'s belief that Bush was elected.

Let's begin with generation of information. An event's occurrence generates information. How much is generated is a function of how likely was the event's occurrence. The more likely an event, the less information it generates – while the less likely the event, the more information it generates. Different ways of classifying events may result in different amounts of information generated. And there are many different ways of trying to measure or quantify amounts of information. Dretske follows the communication industry standard (Weaver & Shannon 1949) of measuring information in bits (*bi*nary digi*ts*), representing the number of binary partitions necessary to reduce a collection of equally probable outcomes to one (e.g., beginning with 8, a three-step reduction to 4, to 2, to 1 = 3 bits). The amount of information generated at a source *s* by the reduction of *n* equally likely possibilities to one is represented: $I(s) = \log n$ (base 2). Here $I(s)$ represents the average amount of information generated at a source by a reduction of equally likely events. If the range of possible events at the source $s_1, s_2, \ldots s_n$, are not all equally likely, then the amount of information generated by the occurrence of s_i is: $I(s_i) = \log 1/p(s_i)$ (where p = probability). So, for example, suppose 10 persons apply for a job and nine are from inside the company, one from outside. If s_1 is the selection for the job of someone outside the company, then $I(s_1) = \log 1/.1 = 3.33$ bits of information. For contrast, selection of someone from inside the company, s_2 would generate $1/0.9 = 0.15$ bits of information.

Next, let's consider information flow or transmission. For information at a receiving point *r* to be about a sending point *s*, there must be dependence between the events at *r* upon those at *s*. Suppose at *s* there are 8 candidates equally likely to be selected. A selection of Susan generates 3 bits of information. Suppose at *r* there are eight equally likely names that may be put on the employment forms in the employment office. A selection of "Susan" generates 3 bits of information. But there would also be 3 bits generated if, mistakenly, the name "Tony" were placed on the employment forms. Clearly, though this amount of information is the same it is not the information that Susan was selected. We want the information at *r* to be about the events that transpired at *s*. Letting "$I_s(r)$" represent this information, $I_s(r) = [I(r) - \text{noise}]$. Noise is the amount of information generated at *r* that is independent of what happens at *s* (not about *s*), and when "Tony" is placed on the forms, but Susan was selected, the noise = 3 bits. Thus, no information about *s* arrives at *r*.

Now for our purposes, the import of these formulae for calculating amounts of information is not so much the absolute values of information generated or transmitted by an event, but the conditions necessary for transmission. For most events it would be difficult or impossible to determine the exact probabilities and ranges of possibilities closed off by an event's occurrence. What is important is whether one receives at *r* as much information as is necessary to know what happened at *s* (under a relevant specification). For a signal or message to carry the information that Bush was elected, it must carry as much information as was generated by Bush's election. We know this is more information than that a Republican ran for office, and more than that someone was elected. Calculating exactly how much information is generated by Bush's election is not as important as determining under what conditions the information that does arrive carries the information that Bush was elected. This is what Dretske calls the informational content of a signal.

Informational content: A signal *r* carries the information that *s* is *F* = The conditional probability of *s*'s being *F*, given *r* (and *k*), is 1 (but, given *k* alone, less than 1).

k is a variable that takes into account how what one already knows may influence the informational value of a signal. If one knew nothing, *k* would go to zero. If I know that Vice President Cheney is from Texas or Wyoming, and I learn that he is not from Texas, I thereby have the information that he is from Wyoming. If you hear that he is not from Texas, but don't already know Wyoming is the only other possibility, you do not thereby receive the information that he is from Wyoming.

This account of the informational content of a signal has important virtues. If a signal carries the information that Bush was elected, then since

the conditional probability that Bush was elected, given the signal, is 1, then *Bush was elected*. Hence, the account gives information a connection to truth. Clearly it will also be the case that the signal carries as much information about *s*, $I_s(r)$, as was generated by the fact that Bush was elected. Noise about the fact *that Bush was elected* is zero. Hence, the account gives us a way to understand transmission or flow of information of a specific propositional (factual) content from source to receiver – not just amounts of information.

Finally, we can give an information-theoretic account of knowledge (ITK): *K* knows that *s* is *F* = *K*'s belief that *s* is *F* is caused (or causally sustained) by the information that *s* is *F*. Dretske says that this is intended to account for perceptual knowledge only, that is, perceptually knowing of something *s* that it is *F*. Knowing, by my current visual experiences of my computer, that it is on, would count as perceptual knowledge. And so would my knowing that the coast is clear, by hearing three knocks, on a prearranged signal. However, here my knowledge also involves knowing the prearranged signal. I know that the coast is clear by hearing the three knocks. Clearly there are other forms of knowledge and other ways of knowing and other kinds of things that can be known. Whether an ITK can be adapted to all cases of empirical knowledge is an interesting question still open. It is not the only one.

Interesting Open Questions

Perhaps the first interesting and open question for an ITK is: how many conditional probabilities of 1 exist? If they are required for knowledge and the world does not provide them, then skepticism rules the day. Knowledge, even if it exists, may be scarce. Famously, Descartes may have known he existed because the probability that he did, given that he was conscious of doubting (thinking), was 1. But is the probability zero that I might have a brain tumor causing me to hallucinate typing this passage? Unless it is zero, I am not receiving (by my experiences) the information that I am typing this passage now, and, according to ITK, do not know that I am. Some people do hallucinate from tumors.

The possibility of the brain tumor may seem extremely remote, because it would have to cause such sophisticated hallucinations. But other possibilities are less remote. Do you know where your car is? Is the probability that it is where you parked it, given your memory impressions of having parked it there, 1? Are cars stolen, where you live? Or, do you know the money in your pocket is not counterfeit? Is the probability that it is counterfeit zero?

Suppose I see a person that I take to be Bill Clinton. I believe it is Clinton because the person looks like Clinton (I've seen him on television many times). Do I know that it is Clinton? I do only if the probability that someone in my vicinity would look just like that (where the "that" refers to the way he looks) and not be Bill Clinton is zero. Or, to put it another way, the probability that it is Clinton, given that it looks like him, must be 1. Now suppose that I get no confirmation – no follow-up information or observation. But suppose that it was Clinton. Do I know this? I do only if there are no "dead ringers" for Clinton in my vicinity – no one who might look like him from an angle or side view, or even straight on. Is the world ever that stable, when it comes to the way people look? If it is not, I could never know it was Clinton from such a chance meeting, because I would never receive the information that it was Clinton. So it is an open question whether the world is stable enough to generate this type of information, and if it is, how often – and about how many things: Clinton, parked cars, legal currency?

One open question was whether the world provides determinate information that *p* (conditional probabilities of 1). Another is whether we can know that it does. To know that the world provides conditional probabilities of 1 is to know that we have the information that *p*. To know that we have the information that *p* is to know that we know, on ITK. Now many would maintain that we do not know that we have knowledge, even if we do. And if that is correct, then we do not know that we have conditional probabilities of 1 in the evidence causing or sustaining our beliefs. So unless we know that we have knowledge, we will not know that there are

the conditional probabilities of 1 that knowledge requires. Thus, it also would be an open question whether we know that there is knowledge. These skeptical challenges are faced by this (or any) theory of knowledge. Here they must be left as open questions. ITK tells us what it takes to have knowledge, but not whether we have it.

Next, on an ITK, since one's belief that *p* contains the information that *p* (Barwise & Perry 1983), should not one necessarily be able to convey one's knowledge that *p* to others? There are reasons (Dretske 1982) to say no. Suppose that Al repairs motorcycles for a living, and that he would never mistake a BMW for a Harley Davidson (only BMWs have horizontal pistons). However, Al mistakenly believes both are American made. When Al is working on a Harley, will he know that it is American made? Al knows that it is a Harley by the label on the motorcycle and the repair manual. The conditional probability that it is American made, given that it is a Harley, is 1, and Al believes it is American made because he believes it is a Harley. So Al's belief that the Harley he is repairing is American made is caused and sustained by items carrying the information that it is American made. But in general, when Al thinks something is American made, the probability that it is, given his belief, is not 1 (for he thinks BMWs are American made). Thus, Al's beliefs of the form [motorcycle *x* is American made] do not contain the information that something is an American motorcycle. Now, if Al is on the phone with someone and says only that he is repairing an American motorcycle, the hearer will not receive the information that Al is working on an American motorcycle. For Al would have said this even if he had been working on a BMW.

Immediately following this discussion, we must ask whether it implies that Al may know that the motorcycle he is working on is American made, but be unable to communicate that knowledge? (This will depend on what the other person on the phone knows about Al. For our purposes, let's suppose that the person on the phone knows very little about Al or motorcycles.) Normally we would expect that if someone knows something, and sincerely asserts the truth that is known, then one who hears the assertion and understands it can thereby come to know the thing known by the speaker. The above example calls this into question. For, as we have supposed, the person on the phone hears Al assert that he is working on an American motorcycle, but does not receive the information that Al is working on an American motorcycle. For Al need not be a reliable conduit of information. Al's utterance may contain only the information that the motorcycle is American made or German made (but not which). Therefore, Al seems to know something that he cannot transmit or pass on – at least if all that he tells the hearer is "it is American made."

It is another interesting and important consequence of this type of ITK that logical, mathematical, or analytic truths generate zero information. Metaphysical necessities (water's being H_2O or Tully's being Cicero) also generate zero information. Since $I(s_i) = \log 1/p(s_i)$, and since any of these things have a probability of 1, their informational value is zero. Hence, we have a feature that puzzled Frege, viz. both that $a = a$, and that $a = b$ generate no information. Yet the latter seems informative, while the former does not. Now it may seem strange indeed to say that it takes no information to know these. And it certainly seems strange to treat nomic necessities (water's expanding when freezing) in the same way one treats mathematical or logical truths.

Perhaps just as puzzling are nomological impossibilities and logical falsehoods. Since they have a probability of zero, the amount of information carried by sentences about them is infinite (or undefined). There are interesting ways out of this conclusion (Floridi forthcoming *a*), but we will not have space to pursue them here.

It is certainly an important and open question whether an ITK can be developed to explain how we know mathematical or logical truths, since these truths generate no information. It is possible that, while no information is generated by these truths, because of their impossibility of being false, they are known by means other than receiving information. Several possible ways this might go are available, but we will not have time to pursue them here. It will have to suffice to say that it might be possible to know things about necessary truths by getting elements of one's beliefs and representational structures

(mathematical and logical representations and proofs) to mirror or be isomorphic to the elements of the domain to be known (Katz 1999). Whether this is possible is still an important and fascinating open question in philosophy of mathematics and theory of knowledge.

Current Philosophical Debates

Is an ITK able to withstand scrutiny? Let's look at some objections. Foley (1987) says that any informational account that relies upon causation of a belief by information will be susceptible to well-known problems of causal deviance. Ironically, this is exactly one of the things an informational account was designed to avoid with the notion of information, because an information channel must screen off causal deviance. So what is Foley's example and does it work? Foley focuses on Dretske's example of three quick knocks at the door causing a spy to believe that the courier has arrived and carrying that information, as well. Foley modifies the case so that it involves a wayward causal chain, as follows. The spy suddenly goes deaf. Then come three knocks carrying the information about the courier's arrival. The knocks cause the spy's partner to trip, causing a box to fall on the spy's head, in turn jarring the spy's brain in such a way that he suddenly comes to believe that the courier has arrived.

As we've seen, ITKs admit that a belief that *p* can be knowledge, even if the belief that *p* does not contain the information that *p*, as long as it was caused by the information that *p*. This invites the kind of example Foley gives. Still, it is quite clear that an information-theoretic account needs the *proximate* cause of the belief (or sustaining cause) to carry the information that *p*. In Foley's example, that is not the case. Even if the three knocks contained the information that *p* and for some strange reason the spy's partner would not have tripped unless the knocks did contain that information, the rest of the story doesn't preserve information. The communication channel has been broken by the time the box falls on the spy's head and his brain is jarred. The conditional probability that the courier has arrived, given the knocks, is 1, but the conditional probability that the courier has arrived, given that the events in the jarred brain have occurred, is not 1. This is because the spy's brain, since jarred sufficiently hard by the box, might be in that state (seeming to hear three knocks) even if the courier had not arrived and there had not been three knocks at the door. Let me explain.

For the purpose of addressing Foley's example and issues to follow, it is important to discuss the matter of an information channel. An information channel condition is any fixed condition (other than conditions existing at the source *s* or receiver *r*) which, by variation of its value, would be able to introduce noise between source and receiver. In order for the information that *a* is *F* to flow from source *s* to receiver *r*, there must be as much information about *a*'s being *F* arriving at *r* as is generated by its occurrence at *s*. That will be possible only if the channel conditions that permit information to flow remain fixed. They must not themselves vary, thereby generating information or nonredundant information. Provided that the channel conditions remain fixed, they do not generate information in the form of noise on the channel, with respect to the information that *a* is *F*.

Let's consider an example. Suppose Al has a metal detector that emits a tone when metal is within 10 inches of its detection surface. For its tone to carry the information that there is metal present, the detector depends on several channel conditions. The power supply must be adequate and charged. The magnetic field that detects the metal must be in place. The wires that activate the tone must be well functioning (no shorts or breaks in the wire). And so on. Were any of these conditions not to remain fixed (a short circuit, say), the detector may emit a tone when there is no metal being detected (even if there is metal present). Therefore, the detector's tone will carry the information that there is metal present only when all channel conditions are *fixed*. In virtue of these fixed conditions, information can flow from the source (detection of metal in the magnetic field of the surface) to the receiver (Al hears the emitted tone).

Notice that the tone carries information about the presence of the metal, but it does not also

carry information about the channel conditions of the metal detector. We can use the tone to tell us about the channel conditions, as we will now see. But we cannot use it to tell us about its channel conditions *and* about presence of metal *at the same time*. If we know in advance that there is metal (or is none), we may check to see whether the detector emits a tone (or does not). Thereby, we can use old information about what we already know (there is/is not metal present) to gain information about the channel conditions of the detector. Now, because we *already know* whether or not there is metal present, we can test to see whether or not the detector is working properly. We can then tell that it is working properly and its channel conditions indeed are fixed (or tell that it is broken because its channel conditions are variable). But if we do not already know in advance about the presence of the metal, the tone carries no (new) information about the channel conditions themselves.

This is not to say that channel conditions last forever. Metal detectors break or wear out. But the point is that for the detector to be a source of information about the presence or absence of metal, its channel conditions must be fixed. When are they fixed? When they generate no (or no new) information.

It is now easy to see that Foley's example is one where noise is introduced into the communication channel because the channel conditions of the man's jarred brain are not fixed. The internal workings of such a brain introduce variability and noise to the system. The same would be true if there were a short in our metal detector (when there was metal present). In neither case (hearing a tone or seeming to hear three knocks) would the relevant piece of information be carried by the tone (or the auditory experience as of three knocks).

Similar remarks apply to an example by Plantinga. Suppose *K* suffers from a brain lesion that causes *K* to believe a variety of mostly false propositions. It also causes *K* to believe that he has a brain lesion, but *K* has no evidence for this belief. Nonetheless, referring to Dretske's ITK analysis above, Plantinga maintains that the "probability on *k* & *K is suffering from a brain lesion* is 1" (Plantinga 1993: 195). Notice that this is not a case of perceptual knowledge. There is no signal that *K* perceives and which carries the information that he has a brain lesion. But even setting this aside, it is clear that the brain lesion is not a fixed-channel condition. It introduces noise to *K*'s cognitive system. Indeed, Plantinga grants that the lesion causes *K* to believe mostly false propositions (a sure sign of noise). Therefore, despite Plantinga's claim, it simply is false that the conditional probability that *K* has a brain lesion, given that *K* believes he has a lesion, is 1. *K*'s beliefs do not track the truth, as Nozick would say. *K's suffering from a lesion* guarantees that he has a lesion, but not his *belief that he has a lesion*, not even if his belief is caused by the lesion (as for a tone from a metal detector with a short caused by its detection of metal). Therefore, the example fails.

However, Plantinga could fix the example so that there is no noise on the channel. So we should consider what ITKs should say to such a case. Indeed, there is a case by Lehrer that will do (1990: 163–4), his Truetemp case. Suppose that quite without his being aware, Mr. Truetemp has a high-tech, belief-producing thermometeric chip implanted in his brain. The chip measures the surrounding temperature and then directly enters a belief that it is that temperature into Truetemp's belief box, if you will. Lehrer and Plantinga would agree that Truetemp does not know the temperature is 56 degrees Fahrenheit, when it is 56 degrees Fahrenheit, and Truetemp believes that it is. As before, there is no signal that Truetemp perceives and which carries the information. The information is mainlined into Truetemp's belief box, as it were. Suppose Truetemp is cognitively unable to withhold belief despite having no evidence that his beliefs are correct.

Now, unlike in Plantinga's example, there do seem to be fixed-channel conditions at work here. The thermometric-cognition mechanism is a fixed condition (let's suppose that it has been in place for years). Since fixed, that it is working reliably generates no (or no new) information. It takes variation in the source (ambient temperature) and faithfully delivers information about it to the receiver (the belief box). It is an information-delivery system no less faithful than the delivery systems of our senses. The probability that ambient temperature is 56 degrees Fahrenheit,

given Truetemp's belief that it is, is 1. The main difference from our senses is that it skips the step of causing a conscious sensation that then causes a reliable belief about what is sensed. The device in Truetemp skips the conscious sensory step, but its delivery of information is, we are presuming, as stable and faithful as any normal sensory perception by standard perceivers under standard conditions.

It would certainly be possible for an ITK to bite the bullet and insist that this is a case of knowledge, albeit a very unusual kind of case indeed. All that Plantinga or Lehrer have to fight this stand is a conflicting intuition and a conflicting theory. Since one probably should place little stock in the conflict of intuition, in the end all they may have is a conflicting theory. Which theory is correct will have to be determined by looking at overall consistency and explanatory power. Saying that this is indeed a case of knowledge may be true and prove not to be so strange, in the long run. It may come to grow on one. Only time will tell.

As with other externalistic accounts of knowledge (Nozick 1981), ITK also rejects the following closure principle for knowledge: $[(K(P) \ \& \ K(P \rightarrow Q)) \rightarrow K(Q)]$ (To be read: knowing that p and knowing that p implies q, entails knowing that q.) Many find this rejection intolerable. I will not here go into the arguments for or against the rejection of this principle (De Rose & Warfield 1999). However, some have pointed out that theories such as ITK will also have to accept a particularly nasty consequence of the rejection of closure, a consequence that is even more intolerable. Following an unpublished example by Kripke, Lehrer (1990) offers the following example designed to show that one might know $P\&Q$, but not know Q (call this a failure of conjunctive closure). Suppose that a Hollywood movie-set has brought fake barns into your countryside. They have erected fake barns of several colors, but not red ones. Indeed suppose it is not possible to fake a red barn. Now if Al believes that something is a red barn ($P\&Q$), he should know because the probability that it is a red barn given that it looks like a red barn is 1. But Al should not know that it is a barn (Q) because the probability that it is a barn, given that it looks like a barn, is not 1 (because the fakes look like barns). So, it seems Al knows $P\&Q$ but not Q – intolerable.

But is this nasty consequence really forced upon an ITK theorist by this type of example? It is not. First, note that this example has the same structure as the case above where Al thinks both BMWs and Harleys are American made. Al knows that the motorcycle he is working on is a Harley (P) and American made (Q). Does he not know that it is American made (Q)? It might be thought that he doesn't because he mistakenly believes some BMWs to be American made. But since he confuses no Harleys with BMWs, he does not believe the Harley is American made because he confuses Harleys with BMWs. He believes it is American made in part BECAUSE he believes it is a Harley. So ITK would say that Al knows it is American because the information that it is American is contained in the information that it is a Harley and Al knows he is working on a Harley.

The Kripke example is no different. If Al believes "this is a barn" (while standing before a red barn illuminated by white light), does he not know that it is a barn? ITKs submit that he does because he *is* receiving the information that it is a barn (contained in the information that it is a red barn). This information is contained in his visual percept. Nothing anywhere near Al looks like a red barn, unless it is both red and a barn. So his percept carries the information that it is a red barn and this, in part, causes his belief that it is a barn (and his belief that it is red). Thus, this particular example is not a case where conjunctive closure of knowledge fails.

Of course, whether it is possible for conjunctive closure of knowledge to fail may still be open. Such failure is certainly consistent with ITKs rejection of closure, generally. There may well be cases of the following sort. One might believe that something is a red barn because of one piece of evidence and continue to believe it is a barn by another piece of evidence. The first piece of evidence might contain the information that this thing looks red and looks to be a barn. The latter piece of evidence may contain only the information that it looks to be a barn. The first contains the information that it is a barn. The latter carries only the information that it is a barn or a fake (but not which). So conjunctive

closure might fail under this type of condition, but *not* under the type of conditions given in the examples by Kripke or Lehrer. Therefore, ITKs are not forced to accept the particularly nasty consequences of denial of closure charged by Kripke and Lehrer.

Interestingly, if it is true that Al knows the structure is both red and a barn (and that the motorcycle is both a Harley and American), then such cases would demonstrate the falsity of theories like Lehrer's (1990: 184). For Lehrer argues convincingly that his theory has as a consequence that Al does *not know* it is a red barn or that it is a barn. If Al knows both, as I believe to be the case, then Lehrer's account of knowledge would be false. As philosophers are fond of saying, one person uses *modus ponens* where another uses *modus tollens*.

References

Adams, F. 1986. "The function of epistemic justification." *Canadian Journal of Philosophy* 16: 465–492. [This paper argues that if epistemic justification permits justified false belief, then it is neither sufficient nor necessary for knowledge. For mature students.]

Barwise, J. and Perry, J. 1983. *Situations and Attitudes*. Cambridge, MA: MIT/Bradford. [An important systematic treatment of the logic and semantics of sentences about language, meaning, belief, and knowledge. Not for the beginner.]

DeRose, K. and Warfield, T., eds. 1999. *Skepticism: A Contemporary Reader*. Oxford: Oxford University Press. [Excellent collection of 15 papers on skepticism, knowledge and closure, and knowledge and relativism. Accessible.]

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT/Bradford. [A seminal attempt to apply developments of the mathematical theory of communication to philosophical issues of knowledge, belief, concepts, and meaning. Accessible to a beginner.]

——. 1982. A Cognitive Cul-de-sac, *Mind*, 91, 109–111. [Argues that avenues of knowledge may dead-end in some knowers – *s* may knows that p, but be unable to transmit that knowledge to another. Accessible to the beginner.]

Floridi, L. Forthcoming *a*. "Outline of a theory of strongly semantic information." *Minds & Machines*. [Explores avenues to escape the Bar-Hillel/Carnap paradox that contradictions or logical impossibilities contain more information than the most important scientific discoveries. Not for the beginner.]

——. Forthcoming *b*. "Is information meaningful data?" *Philosophy and Phenomenological Research*. [Revises the definition of semantic information by showing that false information is not a type of information, and that information is well-formed, meaningful and truthful data. Not for the beginner.]

Foley, R. 1987. "Dretske's 'information-theoretic' account of knowledge." *Synthese* 70: 150–84. [Commentary on Dretske 1981. Accessible.]

Gettier, E. 1963. "Is justified true belief knowledge?" *Analysis* 23: 121–3. [Easily accessible. Perhaps the most widely read paper on knowledge of the twentieth century. The answer is "no," by the way.]

Katz, J. 1998. *Realistic Rationalism*. Cambridge, MA: MIT Press. [Excellent book that attempts to demonstrate the possibility to know mathematical and logical truths despite not coming into causal contact with the objects known. Pitched at the mature student.]

Lehrer, K. 1990. *Theory of Knowledge*. Boulder, CO: Westview Press. [Overview of the problems and topics of knowledge, and presents Lehrer's own systematic view. Accessible, but rigorous.]

Lewis, D. 1996. "Elusive knowledge." *Australasian Journal of Philosophy* 74: 549–67. [Presents and defends Lewis's own view of knowledge, in light of the relevant literature since 1963. Not for the beginner in philosophy.]

Nozick, R. 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press. [First-rate book on many problems of philosophy, including Nozick's own very influential theory of knowledge. Accessible to the student with some appreciation of issues in symbolic logic and counterfactuals.]

Plantinga, A. 1993. *Warrant: The Current Debate*. Oxford: Oxford University Press. [Plantinga's history of post-Gettier epistemology about warrant – i.e., what it is that when added to justified true belief yields knowledge.]

Weaver, W. and Shannon, C. 1949. *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press. [A classic mathematical treatment of information and communication that has been widely influential in philosophy and cognitive science.]

Chapter 18

The Philosophy of Computer Languages

Graham White

1 Introduction: Two Semantic Projects

> Consider a theory whose aim is to say, for a given language, what each of its expressions means. Call it a semantics of that language. What might be required of such a theory before it was allowed that it accomplished its aim or did so in an optimal way? (Travis 1986)

This is a question which belongs to a well-established philosophical program: the Davidson–Dummett program of using a formal semantics – a "theory of meaning," as these philosophers call it – in order to attack philosophical questions about the relation between language and reality, or between mind and language (see, for example, Wiggins 1997).

There is also, among theoretical computer scientists, a similar area of study: it is usually called "the semantics of computer languages," or often simply "semantics." Its aim is, likewise, to develop a formal account of the meaning of a given computer language, and to use that account to answer interesting questions.

These questions might be severely practical – one might, for example, want to formally verify, using such a semantics, that the language in question did what it was claimed to do. However, one might also want to answer more general questions: one might want to design a computer language, and an intended semantics for that language is often a good place to start. Or one might want to classify existing computer languages: it is clear that some of them are more similar to each other than to others, and that some languages are merely notational variants of each other, but how do we put such observations on a more formal footing? And, finally, we might be tempted to say something about the nature of computation on the basis of the semantics of the languages in which we express computations.

I would claim that the semantics of computer language has considerable philosophical interest: it has a different motivation to the usual philosophical approach to computation via Turing machines, and, correspondingly, it yields different insights. One of the main difference is that the semantics of programming languages is concerned with the languages in which people actually program, and, particularly, with the languages which have been found to be good to program in; it is thus inescapably connected with the practice of programming. It is emphatically not a discipline which is developed out of some *a priori* notion of computation. And, in fact, it has

connections with areas of philosophy which may seem surprising; Quine's concept of referential transparency, for example, is an important part of the programming-language enterprise. Some of this material is quite technical, and is also generally unfamiliar to philosophers (even to those who know the usual technical repertoire of philosophical logic). I have segregated many of the more technical details into sections of their own, which can, at first reading, be omitted.

2 History

The semantics of programming languages grew up in a particular historical context, and it is worth spending some time describing it: it was developed by a group of philosophically literate mathematicians and computer scientists, and the philosophical influences are quite evident. They are also little known: the history of computing has tended to focus very much on the very early days, and, at that, mostly on the history of hardware, so that the history of these topics is doubly neglected.

2.1 The first programming languages

In the early days of computers, programming was done by directly writing machine instructions: this was difficult and error-prone. Programming languages were invented to allow programmers to write in a more comprehensible form: Fortran dates from 1954, and made it possible to write programs in a notation very like that of standard mathematics (Backus 1981). Although the Fortran designers paid very little attention to theory – Backus, the leader of the project, says "we simply made up the language as we went along" (1981: 30) – both syntax and semantics soon became important factors in the design of programming languages. Algol was designed over the period 1958–60 (Naur 1981; Perlis 1981), Lisp from 1958 to 1962 (McCarthy 1981), and many of the difficulties of developing these languages were due to two factors: it was difficult to define the syntax of a language at all precisely, and the semantics of these languages seemed utterly mysterious. The latter was an extremely serious problem: without some sort of semantics, it was hard to say what counted as a correct implementation of these languages. Although Lisp eventually achieved a precise semantics, it was designed by starting from the implementation and then attempting to find mathematical structure in the resulting language: several of the Lisp primitives were called after hardware features of the machine that it was originally implemented on (McCarthy 1981: 175), whereas its original semantics was "ramshackle" (Landin 2000). Nevertheless, it was also possible to see that the rewards for a precise syntax – or, more ambitiously, a precise semantics – were extremely high: Algol "proved to be an object of stunning beauty" (Perlis 1981: 88).

2.2 Algol-like languages

What, then, do these programming languages look like? We will describe a generic language, quite similar to Algol; since Algol has had an enormous influence on language design its features can be found in many others. These languages have some similarity to formal languages like first-order logic: like these languages, they have variables, to which values can be assigned, and they have both predicates and functions. And many of the basic operations of programming can be viewed as the assignment of values to variables, so one might think that these operations, too, could be viewed in this way.

However, programming languages also have features which are strikingly different from the sort of logical languages familiar to philosophers. In large part, these other features come from a need to control the structure of programs: programs are extraordinary large entities, and programmers can only keep control of this complexity by making programs out of smaller components, which can be individually constructed and tested and, if possible, reused in many different programs. Programs, then, tend to be made of hierarchically nested components; there are various names for these components, but we can – following Algol usage – call them *blocks*.

Furthermore, unlike the logical languages which philosophers are familiar with – namely,

variants of untyped first-order logic – modern programming languages are typed: variables, and the values that they take on, have types. This is partly for practical reasons: many programming errors can be detected automatically, simply by checking the types of the entities involved. But there are also rather deeper reasons: programming does not fit very well into a set-theoretic view of things, since sets – even finite sets – are collections without any extra structure, and, however we may choose to store data in a computer, we always do so in some structured way (the data may be ordered, or arranged on the leaves of a tree, or the items may be mapped to integers, and so on).

There is a final difference, which is extremely far-reaching. Most programming languages allow programs to perform actions that change the values of variables, or which have other irreversible effects (input or output, for example); we say that these actions have side-effects. These features add further complications to the task of giving semantics to these programming languages.

2.3 The development of semantics

The first steps towards the semantics of these languages were taken, in the 1960s, by a group – Peter Landin, Dana Scott, and others – associated with Christopher Strachey (Scott 1977; Landin 2000). They provided what is called a denotational semantics: that is, rather than describe the operations that pieces of code perform, they associated mathematical objects – denotations, or semantic values – to the syntactic entities of a programming language. Values are assigned to entities on all scales: the variables (and constants) of the language get values, of course, but so do assignment statements – the parts of programs which give values to variables – as do blocks and subroutines and, finally, the entire program. This requirement – that programs should have semantic values on all scales – is part of the basic program of denotational semantics: it is motivated by the view that programming constructs which have well-defined semantic values will be easy to reason about, and it has, over the years, shown itself to be quite justified.

This requirement means that the semantic values assigned to these entities must belong to a quite intricate system. To see this, consider a particular case of programming entities: namely, those that are sometimes called *subroutines*. These are pieces of code that have parameters: they are invoked with particular values of their parameters, and they then perform various actions on them. A subroutine, then, can be thought of as a sort of function: its arguments are the semantic values of its parameters, and its value is the semantic value of the expression that it returns. So the semantic value of a subroutine must be a function type: it maps its argument types to its return type.

But now consider a subroutine which takes a subroutine as a parameter. Such things occur frequently in the normal practice of programming: for example, we might want to write a subroutine which constructed a button in a user interface. Buttons can perform various actions, and – because we are writing a subroutine which we can use for constructing all sorts of buttons – we want to be able to give the code for the action to the button code as a parameter. The "code for the action," of course, is itself a subroutine: so the code for the button is a subroutine which has a subroutine as one of its parameters. The need for "higher-order" entities of this sort seems to be natural and pervasive in programming: our example is from user-interface programming, but there is a differently motivated example in Abelson et al. 1996: 21–31.

A subroutine with a subroutine as argument, then, can be considered as a piece of code which takes the subroutine and returns a value: so its semantic value will map the semantic value of its subroutine parameter to the semantic value of its result. The semantic value of subroutine-calling code, then, is a function which takes another function as an argument: in technical terms, it is a functional.

We must also remember that the functions corresponding to subroutines cannot, in general, be everywhere defined. We know, from the theory of recursive functions, that any reasonably expressive programming language must have programming constructs – looping, recursion, or both – which cannot be guaranteed to give a result in all cases. This must be accommodated in the semantic values of such subroutines.

2.3.1 Technical interlude*: semantic values in detail*

The initial stages of this accommodation are quite easy to see: we can deal with partial functions from, let us say, the integers to the integers by regarding them as functions from **int** – the usual integers – to $\mathbf{int}_\perp$ – the integers together with an extra element, $\perp$, which is the value of the function when the computation fails to terminate. But now a subroutine which takes such a subroutine as argument must itself have a type which is not merely the type of functions (**int** → **int**) → **int**, but rather the type of functions (**int** → $\mathbf{int}_\perp$) → $\mathbf{int}_\perp$ (which is, one should point out, much more complex than ((**int** → **int**) → **int**)$_\perp$). The moral is clear: although we only have to add a single extra element to our base types, the changes required at higher types become progressively more complex.

Recursion, also, needs a suitable treatment. Consider a recursively defined subroutine, for example:

```
function f(x:int): int
begin
  if (x = 0) then f := 1 else
  (f := x * f(x – 1))
end
```

We can regard this as saying that the subroutine f is a fixed point of a certain operation, namely the operation which takes a subroutine F as input and returns, as output, the subroutine G, defined by

```
function G(x:int): int
begin
  if (x = 0) then G :=1 else
  (G := x * F(x – 1))
end
```

So the semantic value of f must be fixed under the semantic counterpart of the operation F ↦ G; and thus, to handle recursion, the appropriate semantic domains must be closed under certain fixed-point operations. Finally, the need to accommodate assignment statements brings another complication. Consider the following subroutine:

```
function f(x: int): int
begin
  y := x;
  f := y + 1
end
```

which first assigns the value of its argument to a global variable y, and then returns y + 1. Consider also the simpler subroutine

```
function g(x: int): int
begin
  g := x + 1;
end
```

f and g yield the same values for the same arguments, but they are not substitutable, one for the other: g changes the values of a global variable, which f does not. (We say that f has *side-effects.*)

Accommodating this sort of behavior, and still preserving the compositional nature of our semantics, makes the type system for our semantic values somewhat complex and intricate – see Tennent 1994: 250ff for details.

The development of semantics, then, is a process of progressive elaboration of semantic values. We might think of it like this: originally, we have a straightforward conception of what the values of programming entities are: variables stand for their values, subroutines stand for functions from parameters to return values, and so on. We may call this original conception the intended semantics. However, it proves impossible to preserve substitutivity with this intended semantics, so we have to progressively elaborate the semantic values that we assign to programs and their parts; in the process, these semantic values become further and further removed from the original, intended semantics.

The divergences are caused by phenomena that can be viewed, when measured against the intended values, as a lack of referential transparency. The intended values of subroutines such as these ought to be given by the maps, from arguments to return values, that they induce, but, as we have seen, subroutines with the same intended values might not be intersubstitutable: and such a failure of substitutivity is, in Quinean terms, described as referential opacity.

3 The Uses of Semantics

A working semantics on these lines can, indeed, be achieved (Stoy 1977; Tennent 1994), and such semantic accounts of programming language have been widely used.

However, the uses are not as direct as one might imagine. It is rarely expedient, for example, to establish correctness for a particular program by examining the semantic values of it and its components: these semantic values are usually extremely complex. They must necessarily be complex: since it is possible to decide whether a given program terminates or not, purely on the basis of its semantic value, there must be facts about the semantic values of programs that are as difficult to establish as the halting problem (i.e. undecidable).

On the other hand, semantics has a large number of metatheoretical uses. One can, for example, establish equivalences between programs, and, more generally, one can develop, and semantically justify, logics (the so-called Floyd–Hoare logics) for reasoning about programs (Tennent 1994: 196ff; Jones 1992); and, unlike direct reasoning with semantic values, these logics are, for typical problems, easy to work with.

There is another, less formal but very pervasive, use of semantics. The practice of programming involves a great deal of substitution: replacement of one subroutine by another (or one object by another, one library by another, and so on). We like to have languages in which substitutions like this are easy to justify: if we can be sure that, if two items "behave the same" (in some suitably informal sense) they can safely be substituted for each other. In the development of semantics, it very soon became apparent that the semantic properties of languages were decisive for this question: that certain semantic properties made substitution behave well.

3.0.1 Technical interlude*: an issue in programming-language design*

Here is an example of the sort of guidance that semantics can give in language design. Suppose we define a subroutine – call it S, and that we later invoke it. Suppose also that, in the code defining S, there is a global variable x. Suppose, finally, that we change the definition of x between the time that S is defined and the time that it is invoked. Which value do we use for x? There are two obvious choices:

1. the value it had when S was defined: this is called *lexical binding*, and
2. the value it had when S was invoked: this is called *dynamic binding*.

It turns out (Stoy 1977: 46ff) that lexical binding gives a language much better substitution properties, and, in fact, languages with dynamic binding – the typesetting language TeX, for example – are terribly difficult to program with. More generally, it seems to be the case that, if a language has clean, elegant semantic properties, then it will be easy to program in.

3.1 Identity of programs

We use semantics, then, to conclude facts about the behavior of programs on the basis of mathematical properties of their semantic values. We could, for example, observe that, if the semantic values of programs P and Q were different, then the programs themselves must be different.

This is more subtle than it might seem. What do we mean by identity and difference between programs? A trivial answer would be that it simply consisted in the identity or difference of their source code: but this is rarely of any interest. Programs can vary a good deal, in a merely notational way, and still remain "essentially the same" (whatever that might mean).

A better criterion for the identity of programs is that of *observational equivalence*; philosophically, it can be regarded as a sort of functionalism (see Lycan 1995, Block 1995). One definition is as follows: Two programs, P and Q, are observationally equivalent if and only if, whenever the inputs of P and Q are the same, then so are their outputs.

We can define this also for constituents of programs (subroutines, statements, blocks, and the like: these are generically called *program*

phrases). We define equivalence by observing what happens when phrases are substituted for each other in programs (here a program with a phrase deleted is called a *program context*):

> Two program phrases, f and g, are observationally equivalent if and only if, for any program context $P(\cdot)$, the two programs $P(f)$ and $P(g)$ are observationally equivalent.

Observational equivalence is an extremely versatile property. If we think of anything which might be (in the nontechnical sense) an "observation" of the behavior of a program – stimulating it with certain input, checking the output, and so on – we can automate this observation by writing another program to perform it. This other program will provide a program context with which we can test the program that we are interested in: and thus our definition of observational equivalence can be regarded as an automated version of the everyday concept of observation.

However, observational equivalence is a difficult property to establish, since it talks about what happens when a program fragment is substituted into any program context at all, and in most cases we have no grasp of this totality of program concepts.

If our semantics respects observational equivalence, then we call it *fully abstract*:

> A semantic valuation $\upsilon(\cdot)$ is fully abstract if, whenever program phrases f and g are observationally equivalent, $\upsilon(f)$ and $\upsilon(g)$ are the same.

Full abstraction is, of course, an important concept, because we are interested in observational equivalence. But its interest is rather wider than that: a fully abstract semantics will, in some way, reflect the essential structure of programs, abstracting away from notational or implementational details (Tennent 1994: 242). Of course, we can – rather fraudulently – define fully abstract semantics by starting with a non-fully abstract semantics and imposing equivalence relations on it; but unless we have independent access to the model thus constructed, it would do us no good. In the case when we can find a fully abstract model, and characterize it in some meaningful way – for example, in terms of games (Abramsky et al. 1994; Hyland & Ong 2000) – we have a mathematical object which tells us a great deal about the deep structure of a particular programming language.

3.2 Functional programming

We have been describing an approach to the theory of programming languages which simply seeks to analyze the usual languages that people program in. We might, though, take a different approach to language design: we might want the theory to be more prescriptive, and design programming languages so that they had a good, perspicuous, metatheory.

One of the features which give languages a good metatheory is referential transparency: the property that terms of the language, which stand for the same entities, can always be substituted for each other. Languages with side-effects – such as statements that change the values of variables – do not have referential transparency. A term of such a language might stand for, let us say, a number, but might also, in the course of evaluating that number, change the values of a particular variable; another term might evaluate to the same number, and might change the values of other variables; and it is easy to see that these two terms, even though they evaluated to the same numbers, could not be substituted for each other.

So we might consider designing a programming language in which we could not change the values of variables. What would such a language look like? There could be variables, and we could have definitions that assigned values to variables: however, once we have let a variable have a certain value, we could not subsequently change it. We could also have subroutines: subroutines would take parameters and return values. Because we have no assignment statements, the result returned by a subroutine on particular arguments depends only on the values of its arguments: the same subroutine, evaluated on the same argument, always yields the same result. Subroutines, then, are extensional: they give the same results on arguments with the same referents, and in this respect they behave like mathematical functions.

Following Quine, it is usual to refer to languages with this property as *referentially transparent*.

As we have seen, programming needs higher-order constructs. These languages are no exception: we can let higher-order entities (in this case, functions and higher-order functionals) be the values of variables, we can pass them as parameters to subroutines, and so on. Following Quine's slogan "to be is to be the value of a variable" – a slogan explicitly used by the pioneers of programming-language semantics – we can, and do, give a rough ontology to our language: the entities that can be the values of variables, that are passed to and returned by subroutines, are usually known as "first-class citizens," and they will figure largely in any account of the semantics of the language. These languages – known as functional languages – generally have a large array of such higher-order constructs. Lisp is such a language, but is semantically somewhat impure: modern, more principled versions are untyped languages such as Scheme, and typed languages such as ML.

3.2.1 Technical interlude*: the lambda calculus*

There is an alternative description of these languages. Consider a subroutine, such as the following:

```
begin function f(x, y):
   return 3 * x + 2 * y
end
```

This is the subroutine which takes two parameters, x and y, and returns 3x + 2y. We can give an alternative description of this – in more mathematical notation – as a term in the λ-calculus:

$$\lambda x.\lambda y.(3x + 2y)$$

This rough analogy can be made precise: we can set up a correspondence between functional programs and λ-calculus terms, which is compositional and extensional, in such a way that we can obtain a semantics for our programs from a semantics for the λ-calculus.

We can go on from here. It is known that terms in a suitable λ-calculus can be used to encode proofs in higher-order intuitionistic logic (Lambek & Scott 1986); this is called the *Curry–Howard correspondence*. This suggests that functional programs can also be considered to be proofs in that logic: and such in fact is the case. The correspondence between programs and proofs is illuminating in its own right. Consider a subroutine which takes a parameter – say x – and which computes, for example, $3x + 2$. This corresponds to the lambda-term

$$\lambda x : \mathbf{int}.3x + 2,$$

which corresponds to a proof of the proposition

$$\forall x : \mathbf{int} \exists y : \mathbf{int}$$

but not, of course, just *any* proof: it is the proof which takes the integer x introduced by ∀, computes $3x + 2$, and then uses *that* integer as a premise in ∃-introduction.

4 Conclusions

We have surveyed a rather large area of theoretical computer science; we now have to consider its philosophical relevance. There are, of course, several points of direct relevance: programming-language semantics tells us a great deal about the processes of abstraction involved in programming computers, and also about the nature of algorithms (for which Moschovakis 2001 is a good comparison). However, there are less direct points of interest. The overall goal of programming semantics is quite similar to the philosophical project of developing a theory of meaning: however, the methods and results are strikingly difficult. To a large extent this is because the philosophical project has been developed in isolation, with unsophisticated technical tools, and with the aid of a very small number of examples, none of them either large or complex. The practical necessities of producing a useful body of theory have made it impossible for programming-language semantics to indulge in any of these luxuries. So, the comparative use

of programming-language semantics is, perhaps, more interesting than its direct use.

4.1 What aren't we interested in

Programming-language semantics was developed in order to address certain specific needs, and one must understand the biases resulting from those needs to be able to understand the theory and its place in the world.

We know the mechanism The first is obvious: we know all about the mechanisms of computers, because, after all, we made them. This contrasts very strongly with the situation in the philosophy of language and in linguistics, where we have very little information about underlying neurophysiological mechanisms.

We design the systems We also design computers, their operating systems, and the programming languages that we use on them. Many of the choices that we make when we do this are reflections of our needs, rather than of the nature of computation as such; for example, operating systems are extremely modular, and most programming languages have a great deal of support for modularity, simply because modularity makes computers much easier to program. Modularity does not seem to be entailed by the nature of computation as such. By contrast, Fodor's work (Fodor 1983) deploys much more transcendental premises: he is attempting to establish that any minds such as ours must be modular, whatever their mechanisms and however those mechanisms might have arisen. The semantics of programming languages can, of course, neither prove nor disprove the validity of a program like Fodor's – though, of course, it might provide illuminating results on the nature of modularity and on its formal analysis.

Reference is not problematic If we are using computers to solve a problem in the real world, we generally know what the expressions of our programming language stand for: we have, if we are sensible, set things up that way.

Foundationalism is not interesting We may, in principle, *know* the physical processes that the expressions of our programming languages result in, when they are suitably compiled and run. However, we are very rarely interested: looking at computers on that sort of level would submerge the interesting features in an ocean of low-level detail. We would be unable to distinguish, *on that level*, between operations which were performed by the operating system and those which were performed by programs; of the programs running on a real computer, by far the majority would be concerned with trivial housekeeping tasks, rather than anything we were interested in: of the instructions which execute in an interesting program, by far the majority of them would be concerned with rather dull tasks such as redrawing windows on the screen, interacting with the operating system, and so on. On the level of machine-level instructions, none of these processes would be distinguishable from each other in any tractable way.

4.2 What are we interested in?

So what are the interesting problems?

Making languages different There is a huge number of different programming languages, and there are also genuine differences between them. If we were to analyze these languages using the methods of recursive function theory, or by representing programs written in them as Turing-machine programs, we would find (since programming languages are generally Turing complete) that they were indistinguishable from each other. So we want a theory that is finely grained enough to be able to represent the genuine differences between languages. On the other hand, we do not want to differentiate languages that are merely typographical variants of each other, or which differ simply by trivial definitional extensions: we want, that is, a theory that is sensitive to *genuine* differences between languages, and only to those. We would also like to go on and construct a *taxonomy* of languages: that is, we would like to arrange languages in some sort of formal scheme in which we could describe how to get from one language to another by regularly varying theoretical parameters of some sort.

Attaining abstraction There is a common theme running through all of these considerations. When we are designing, or using, a high-level programming language, we are concerned about attaining a sufficient degree of abstraction. Low-level, detailed, grounded descriptions of our systems are unproblematic, but we do not want these: we want to be able to forget about such merely implementational details in order to program, and reason about programs, at the level we are interested in. In order to do this, we need to be able to design languages to do it; and in order to do that, we need some sort of theoretical conception of what these languages should look like. Thus, our semantics should give us a view of computational processes which is equally as abstract as the languages that we want to design. Abstraction, then, is an achievement.

This contrasts sharply with the traditional task of the philosophy of language. Here we start with a high-level view – a speaker's intuitions about language – and we attempt to find a suitably grounded account of this high-level view (Dummett 1991: 13). Attaining the abstract view is not a problem: grounding it is. In the semantics of computation, on the other hand, we already have a grounded view of our subject-matter: it is the construction of a suitably abstract view that is the major difficulty.

4.3 The technical tools

The technical tools used also differ strongly from those current in the philosophy of language community. Programming-language semanticists use higher-order logic and the mathematical theory of categories: linguistic philosophers use first-order logic and set theory. This is a fairly profound difference in mathematical cultures, but it also has to do with the difference between the problems that these communities are addressing.

Intuitionistic logic Semantics uses intuitionist logic a great deal: we have seen, above, that the semantics of functional programming looks very like the proof theory of higher-order intuitionist logic, because programs correspond to proofs of certain propositions. For this, we must use intuitionist, rather than classical, logic: because we want to make programs correspond to proofs, there must be a large number of essentially different proofs of the same proposition. We can rephrase this in terms of equivalence of proofs: we should be able to define a notion of proof equivalence which disregards merely notational variation, but which is not so coarse that the set of equivalence classes becomes trivial. It turns out that, for technical reasons, we can do this for intuitionist logic, but not for classical logic (Girard 1991).

However, this use of intuitionist logic is less ideological than it might seem. We are using it because we need a finely-grained proof theory, and it would be perfectly possible for an ideologically classical logician to use intuitionist logic for these purposes, only because, for computational purposes, one needed a fine-grained proof theory.

In a similar way, we use higher-order logic: higher-order constructions are pervasive in programming, and it is appropriate to have a metatheory which reflects that. However, this preference, again, is not straightforwardly ideological: these higher-order entities are, after all, algorithms, and carry no taint of the infinite. Correspondingly, there are constructive models of set theory in which sets are modeled by equivalence relations on subsets of the integers: we can, in such models, carry out all of the constructions needed to develop programming-language semantics, although we cannot quite accommodate all of traditional higher-order logic (Robinson 1989; Hyland 1982). There is very little that a constructivist can find about such models to object to.

Category theory is also part of the semantic toolkit (see Tennent 1994: 290ff for some examples). But this is hardly any surprise: category theory has found wide application in areas of mathematics – algebraic topology, algebraic geometry, and proof theory – where one wants to disregard "implementational" (or merely notational) detail, and concentrate on the essential features of a situation.

This can be rephrased in more traditional philosophical terms as follows. There is a well-known example, due to Benacerraf (1965), which is (slightly rephrased) as follows: consider two mathematicians (A and B) who both talk about

ordered pairs, except that A encodes the ordered pair $<x, y>$ as $\{x, \{x, y\}\}$, whereas B encodes it as $\{y, \{x, y\}\}$. Now – though A and B clearly each have accounts of ordered pairs which are mathematically adequate – they do not seem to be talking about the same objects (and, in fact, they can be made to disagree by asking them stupid questions of the form ("is x a member of $<x, y>$?")). One approach to this would be to invoke a difference between specification and implementation, and to say that A and B were simply using different implementations of a single specification. Of course, to do that we need to have some way of making these specifications explicit: and category theory gives us that. We can, given two sets X and Y, specify their Cartesian product $X \times Y$ (the set of ordered pairs with members in each set) in terms of the two maps $X \times Y \rightarrow X$ and $X \times Y \rightarrow Y$, and of the properties of these two maps. And this characterization turns out to characterize the construction exactly, without involving any purely implementational decisions.

We can think of category theory as ruling out the stupid questions which differentiated between A's and B's mathematics: that is, of giving us a distinction between observable and unobservable properties of mathematical constructions. The observable properties are those which can be expressed in terms of mappings ("morphisms," in category-theoretic terms) between mathematical objects, and in terms of identities between those morphisms; the unobservable ones need identities between objects (Bénabou 1985). It is no surprise, then, that programming-language semantics, which is intimately tied to the observable properties of computer programs, also uses category theory to express that notion of observability.

4.4 Theories of meaning

Finally, we should compare these semantic theories with the philosophical project of a theory of meaning. We should recall that the goal of Davidson's program was to develop an axiomatic theory which would yield, for each sentence of a natural language, a suitable instance of the schema (Dummett 1991: 63)

S is true if and only if **A**.

Truth is not particularly salient in the semantics of programming languages, but we do have an important central notion: that of observational equivalence. So, if we do have a fully abstract semantics, then it can (after suitable manipulation to express it in philosopher-friendly terms) be construed as a sort of counterpart of a theory of meaning: it is a mathematical theory from which we can derive a great number of conclusions about observational equivalence of computer languages. And there are such fully abstract semantics.

However, there are one or two caveats to be made. There is a presumption that, when one had achieved a theory of meaning, one could simply examine to see what its "central notion" was (Dummett 1991: 34). But these semantic theories are possibly a little more recalcitrant: they assign mathematical objects – semantic values – to program phrases, but these mathematical objects do not wear their meanings on their sleeve: there is still room for considerable argument about what they mean. We may, it is true, present mathematical objects using vocabulary which is sufficiently tendentious to make one think that they have a clear and obvious meaning: but this would be *merely* tendentious, having to do with a particular presentation of those objects.

The other caveat is this. The semantics of programming languages has paid particular attention to the question of full abstraction: this concept has been somewhat neglected in the philosophy of language, where the problems have seemed to be those of finding rich enough semantic values to hold all of the components of meaning that we want. However, full abstraction ought to play a role in the philosophy of language as well: as Quine said, there should be no identity without identity (Quine 1969: 23), and the semantic values that we assign to sentence fragments should, in some way, respect the identities of the meanings that we are trying to model.

References

Abelson, Harold, Sussman, Gerald Jay, and Sussman, Julie. 1996. *Structure and Interpretation of Computer Programs*, 2nd ed. Cambridge, MA: MIT Press. [An outstanding, and semantically aware,

programming textbook, using the functional language Scheme.]

Abramsky, Samson, Jagadeesan, Radha, and Malacaria, Pasquale. 1994. "Full abstraction for PCF (extended abstract)." In M. Hagiya and J. C. Mitchell, eds., *Theoretical Aspects of Computer Software*. Lecture Notes in Computer Science no. 789. New York: Springer-Verlag.

Backus, John. 1981. "The history of Fortran I, II and III." In R. L. Wexelblat, ed., *History of Programming Languages*. New York: Academic Press. From the ACM SIGPLAN History of Programming Languages Conference, June 1–3, 1978.

Bénabou, Jean. 1985. "Fibred categories and the foundations of naïve category theory." *Journal of Symbolic Logic* 50: 10–37.

Benacerraf, Paul. 1965. "What numbers cannot be." *Philosophical Review* 1974.

Block, Ned. 1995. "Functionalism (2)." In Guttenplan 1995.

Fodor, Jerry A. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

Dummett, Michael. 1991. *The Logical Basis of Metaphysics*. London: Duckworth.

Girard, Jean-Yves. 1991. "A new constructive logic: classical logic." *Mathematical Structures In Computer Science* 1(3): 255–96.

Guttenplan, Samuel, ed. 1995. *A Companion to the Philosophy of Mind*. Oxford: Blackwell.

Hyland, J. M. E. and Ong, C.-H. L. 2000. "On full abstraction for PCF." *Information and Computation* 163: 285–408.

Hyland, Martin. 1982. "The effective topos." In A. S. Troelstra and D. van Dalen, eds., *L. E. J. Brouwer Centenary Symposium*. Studies in Logic and the Foundations of Mathematics no. 110. Amsterdam: North-Holland.

Jones, C. B. 1992. *The Search for Tractable Ways of Reasoning About Programs*. Tech. rept. UMCS-92-4-4. Dept. of Computer Science, University of Manchester.

Lambek, J. and Scott, P. J. 1986. *Introduction to Higher Order Categorical Logic*. Cambridge Studies in Advanced Mathematics no. 7. Cambridge: Cambridge University Press.

Landin, Peter J. 2000. "My years with Strachey." *Higher Order and Symbolic Computation* 13: 75–6.

Lycan, William G. 1995. "Functionalism (1)." In Guttenplan 1995.

McCarthy, John. 1981. "History of Lisp." In R. L. Wexelblat, ed., *History of Programming Languages*. New York: Academic Press. From the ACM SIGPLAN History of Programming Languages Conference, June 1–3, 1978.

Moschovakis, Yiannis N. 2001. What is an algorithm? *Pages 919–936 of*: Engquist, Björn, & Schmid, Wilfried (eds), *Mathematics unlimited – 2001 and Beyond*. Berlin: Springer.

Naur, Peter. 1981. "The European side of the last phase of the development of Algol." In R. L. Wexelblat, ed., *History of Programming Languages*. New York: Academic Press. From the ACM SIGPLAN History of Programming Languages Conference, June 1–3, 1978.

Perlis, Alan J. 1981. "The American side of the development of Algol." In R. L. Wexelblat, ed., *History of Programming Languages*. New York: Academic Press. From the ACM SIGPLAN History of Programming Languages Conference, June 1–3, 1978.

Quine, Willard van Ormond. 1969. "Speaking of objects." In *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Robinson, E. 1989. "How Complete is PER?" In *Proceedings of the Fourth Annual Symposium on Logic in Computer Science*. IEEE Press: Washington.

Scott, Dana S. 1977. "Introduction." In Stoy 1977.

Stoy, Joseph E. 1977. *Denotational Semantics: The Scott–Strachey Approach to Programming Language Theory*. Cambridge, MA: MIT Press.

Tennent, R. D. 1994. "Denotational semantics." S. Abramsky, D. M. Gabbay, and T. S. E. Maibaum, eds., *Handbook of Logic in Computer Science*, vol. 3. Oxford: Oxford University Press.

Travis, Charles. 1986. "Introduction." In C. Travis, ed., *Meaning and Interpretation*. Oxford: Blackwell.

Wiggins, David. 1997. "Meaning and theories of meaning: meaning and truth conditions from Frege's grand design to Davidson's." In B. Hale and C. Wright, eds., *A Companion to the Philosophy of Language*. Oxford: Blackwell.

Chapter 19

Hypertext

Thierry Bardini

Introduction: Defining Hypertext

Defining hypertext can be confusing. Kathleen Gygi (1990: 282) categorized available definitions into two types, "broad-spectrum" (Group I) and the "more clinical variety" (Group II). She found Group I definitions in the popular press and in advertising and marketing literature, and Group II definitions in technical journals and research efforts at developing computer-supported hypertext systems. She gave the following examples:

Group I

- Hypertext works by association rather than indexing.
- Hypertext is a format for nonsequential representation of ideas.
- Hypertext is the abolition of the traditional, linear approach to information display and processing.
- Hypertext is nonlinear and dynamic.
- In hypertext, content is not bound by structure and organization.

Group II

- Hypermedia is a style of building systems for information representation and management around a network of nodes connected together by typed links.
- Hypertext is: (1) a form of electronic document; (2) an approach to information management in which data is stored in a network of nodes and links. It is viewed through interactive browsers and manipulated through a structure editor.
- Hypertext connotes a technique for organizing textual information in a complex, nonlinear way to facilitate the rapid exploration of large bodies of knowledge. Conceptually, a hypertext database may be thought of as a directed graph, where each node of the graph is a (usually short) chunk of text, and where the edges of the graph connect each text chunk to other related text chunks. An interface is provided to permit the user to view the text in such a database, traversing links as desired to explore new areas of interest as they arise, check background information, and so forth.
- Windows on the screen are associated with objects in a database, and links are provided between these objects, both graphically (as labeled tokens) and in the database (as pointers).

More recently, Luciano Floridi (1999) has synthesized these Group II definitions to put the emphasis on the three necessary elements that make a hypertext. He described them as:

1 a discrete set of semantic units (nodes). . . . These units, defined by Roland Barthes as lexia . . . , can be
   1.1 alphanumeric documents (pure *hypertext*)
   1.2 multimedia documents (*hypermedia*)
   1.3 functional units . . . , in which case we have the *multifunctional hypertext* or *hypermedia*
2 a set of associations – links or hyperlinks embedded in nodes by means of special formatted areas, known as source and destination *anchors* – connecting the nodes. These are stable, active cross-references which allow the reader to move immediately to other parts of a hypertext
3 an interactive and dynamic interface.

The emphasis on this third and crucial component is worth noting since it is its presence that dispels two of the six fallacies that Floridi associates with hypertext: the *literary fallacy* according to which "hypertext began primarily as a narrative technique and hence it is essentially a new form of literary style" and the *expressionist fallacy* that holds that "hypertext has arisen as and should be considered primarily a writing-pushed phenomenon" (ibid.). Both fallacies, however, are quite common and in fact define a whole subfield of hypertext theory. George P. Landow might be the most famous representative of such a theoretical insight on hypertext, that focuses on the first two elements of Floridi's standard model . . . and risks forgetting the third: "Hypertext . . . denotes text composed of blocks of text – what Barthes terms a *lexia* – and the electronic links that join them" (Landow 1992: 4). In fact, Landow even claimed that

> The many parallels between computer hypertext and critical theory have many points of interest, the most important of which, perhaps, lies in the fact that critical theory promises to theorize hypertext and hypertext promises to embody and thereby test aspects of theory, particularly those concerning textuality, narrative, and the roles and functions of reader and writer. (ibid.: 3)

Thus, for Landow, the proviso that the links may be "electronic" does not change the basic fact that we are still talking about text, authors, and readers, writing and reading, when we talk about hypertext. It is quite a different position that we will argue for in this chapter, where we will wonder how exactly a set of electronic links between lexia can actually *embody* anything.

One should also note that Floridi's *standard definition of hypertext* takes for granted the equation of *interface* with *electronic interface*: the proviso that the interface should be both *interactive* and *dynamic* makes it almost necessarily electronic, unless one holds a very loose notion of interactivity. Floridi's insistence that hypertext is not *uniquely* a computer-based concept is correct (the electronic fallacy), but his very own standard definition somehow implies that hypertext has been recently redefined to be quasi-uniquely a computer-based *implementation*.

This chapter starts from this premise, that we should consider as an historical fact. Actually, the first section will describe how the development of computer-implemented hypertext since the early 1960s has stemmed from a dual origin that somehow reconciles the literary and the computerized nature of the field. From then, we will proceed with a general discussion on the question of the hypertextual interface, only to get back later to some elements of literary and, better yet, dramatic theory, applied to computer-based hypertext.

Association vs. Connection: The Dual Origins of Hypertext

The standard definition of hypertext is indeed the result of an historical process, in which the organization of its three basic components – a discrete set of lexias, a set of links, and an interface – have been progressively stabilized through negotiations among actors in the field.

The term "hypertext" is usually credited to Ted Nelson, who coined it in 1962 with the idea of hyperspace in his mind. According to him, his influence was mainly found in the vocabulary of mathematics, where the prefix *hyper* means "extended and generalized" (Nelson, personal interview, 3/17/93). To Nelson, hypertext was a necessary tool for his work as an author,

a tool that "allows you to see alternative versions on the same screen on parallel windows and mark side by side what the differences are" (ibid.). From then on, hyperspace slowly became cyberspace.

At the same time that Ted Nelson coined the term hypertext, Douglas Engelbart was beginning to implement his framework for the Augmentation of Human Intellect at Stanford Research Institute (SRI, in Menlo Park, CA). Altough his framework itself did not explicitly mention hypertext, the core of Engelbart's vision was based on a very similar premise, "this extremely flexible way in which computers can represent modules of symbols and can tie them together with any structuring relationship we can conceive of" (Engelbart, personal interview, 12/15/92). The introduction of an hypertext-like capability in Engelbart's framework responded, however, to a very different motivation than Nelson's. It was based on the premise that computers should be able to perform as a powerful auxiliary to human communication and collaboration if they were to manipulate the symbolic language that human beings manipulate.

For Engelbart, what seemed promising about computers was that the processes that match human and machine to the outside world and to each other could be found in natural language, understood in both its physiological and social dimensions; a language available to all. Of all the tools humans use, language clearly seemed the metatool, the one that made all the others possible. What Engelbart meant by "language" was "the way in which the individual parcels out the picture of his world into the concepts that his mind uses to model the world, and the symbols that he attaches to those concepts and uses in consciously manipulating the concepts" (Engelbart 1962: 9). He thus conceived language as operating at two levels: concept structuring but also symbol structuring, in order to model and at the same time to represent "a picture of the world." As Engelbart himself pointed out, this understanding of what language is and does derives from the work of Benjamin Lee Whorf, and it both mirrors and extends the famous Sapir–Whorf Hypothesis: "Both the language used by a culture, and the capability for effective intellectual activity, are directly affected during the evolution by the means by which individuals control the external manipulation of symbols" (ibid.: 24).

Engelbart thus postulated a dialectical relationship between the two sublevels of natural language, a relationship in which the symbolic representation of concepts can affect the way these concepts structure the world. The computer could become an open medium that could be used to "make sense of the world," to map the structure of the world as information flows in order to exploit the potential of natural language to reconfigure our concepts and change our world. The key to this reconfiguration lay not in any single concept itself, but in their being already configured – already given in nonlinear relationships that could be identified, mapped, and changed. When one stretches the notion of technology to include the way humans use language – as Engelbart realized very early, according to his own account – it becomes clearer how it was the influence of Whorf that was central to the development of hypertext.

Because of how he conceived of the way that natural language could function in the human–computer interface, Douglas Engelbart, along with Ted Nelson, often is credited for pioneering work in the field of hypertext or hypermedia. Many, however, trace the genealogy of hypertext not to Engelbart's extension of the Sapir–Whorf Hypothesis, but to the work of Vannevar Bush.

In a famous article entitled "As We May Think," Vannevar Bush proposed a new kind of electro-optical device, the memex, "a enlarged intimate supplement of an individual's memory." The result of "utopian fiction and speculative engineering," the memex was an imaginary machine that existed entirely on paper and that never was constructed (Nyce & Kahn 1991: 45). Bush conceived his memex on the basis of analogies between brain and machine, between electricity and information:

> The human mind operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain . . . Man cannot hope fully to duplicate this

> mental process artificially, but he certainly ought to be able to learn from it. (Bush 1945: 101)

Nyce and Kahn (1991: 60) argue that since Bush's work, "whether access and use of the records should be based on abstract general principles or on personal, i.e., individual associations has been the major issue separating information retrieval systems from hypertext." Hypertext systems in this formulation thus rely on the individual process of "association" as envisioned by Bush, rather than on "abstract general principles," and "for Bush, and later for Nelson and others engaged in hypertext research, memex represented a very personal tool" (ibid.). The term "hypertext" wasn't coined until the early 1960s, however, and Bush himself never used the term to describe his work. The reigning term at the time for what Bush was proposing was indeed "information retrieval." It seems difficult to dispute, therefore, that the memex was not conceived as a medium, only as a personal "tool" for information retrieval. Personal access to information was emphasized over communication. The later research of Ted Nelson on hypertext is indeed very representative of that emphasis.

It is problematic, however, to grant Bush the status of the "unique forefather" of computerized hypertext systems. For the development of hypertext, the important distinction is not between personal access to information and communication, but between different conceptions of what communication could mean, and there were in fact two different approaches to communication at the origin of current hypertext and hypermedia systems. The first is represented by Ted Nelson and his Xanadu Project, which was aiming at facilitating individual literary creativity. The second is represented by Douglas Engelbart and his NLS, as his oN-Line System was called, which was conceived as a way to support group collaboration. The difference in objectives signals the difference in means that characterized the two approaches. The first revolved around the "association" of ideas on the model of how the individual mind is supposed to work. The second revolved around the intersubjective "connection" of words in the systems of natural languages.

What actually differentiates hypertext systems from a information-retrieval systems is not the process of "association," the term Bush proposed as analogous to the way the individual mind works. Instead, what constitutes a hypertext system is the presence of links between lexias on a human–computer interface. And a process of association analogous to the way the individual mind works is not the only way of establishing such links. The most important ones already are established in natural language. Bush himself stated that "The process of tying two items together is the important thing" (Bush 1945: 103). Thus, what actually defines hypertext is the existence of links organizing the information in such paths, regardless of the process by which these links were created.

To put it another way, "association" is only one kind of "connection," and is in fact the least desirable kind, where communication is the goal, precisely because it is the way an *individual* mind works. This distinction too was pointed out by Benjamin Lee Whorf:

> Connection is important from a linguistic standpoint because it is bound up with the communication of ideas. One of the necessary criteria of a connection is that it be intelligible to others, and therefore the individuality of the subject cannot enter to the extent that it does in free association, while a corresponding greater part is played by the stock of conceptions common to people. (Whorf 1927: 36)

Associations can be individual and open-ended. "A common stock of conceptions," by contrast, tends to be limited, and often is structured, even hierarchical, and susceptible to rearrangement. Ted Nelson, who like Engelbart was very familiar with Whorf's writings, stressed that the main difference between his views and Engelbart's view indeed concerned the role of structure and hierarchy: "To me hierarchy is a special case. I don't say that hierarchies are always invalid, it's just that because they're so convenient they've been used too much. And they represent many things very badly" (Nelson, interview, 1993). For Engelbart, with his concern for communication, the opposite was the case.

The Language Machine and the Body

A common ground to all hypertext systems, regardless of their emphasis on hierarchy, is the issue of electronic display/access to linked lexias, often dubbed nonlinear (e.g. Landow 1992: 101–19). But the emphasis on the notion of linearity, again, often reflects back to the model of the text, and might miss the essential third component of the standard model of hypertext, interface. In this sense the interface is indeed the crucial element of an hypertext system, because it is at the interface that the linked lexias are displayed/accessed. Now, what exactly does *interface* mean?

The point of contact between human and computer, the boundary that separates and joins them, usually has been located only via metaphor. To say that "the mind is a meat machine," or, more accurately, that "the mind is a computer," is to use a metaphor: the statement relies on an analogy that "invites the listener to find within the metaphor those aspects that apply, leaving the rest as the false residual, necessary to the essence of the metaphor" (Newell 1991: 160). With the artificial intelligence mind-as-computer metaphor (see Chapter 9, THE PHILOSOPHY OF AI AND ITS CRITIQUE), the greatest source of this false residual lies in the human's direct perception of the computer. Physically, materially, minds and machines or computers are fundamentally different things, however much there may be resemblances that permit metaphorical comparisons.

When one considers the mind-as-computer metaphor as a means to make sense of the "boundary," the obvious conclusion is that the compared materiality of human beings and computers is the false residual of the mind-as-computer metaphor. One should conclude that there is no "natural" way to locate the boundary that distinguishes and joins them. There is no ontological connection, that is, between our materiality – our bodies – and the material manifestation of the computer. But the ultimate goal of the project to create artificial intelligence was to achieve the material realization of the metaphor of the computer as a "colleague," and therefore as a mind, a machine that can pass the Turing Test.

The greatest philosophical achievement of the AI research program might very well be that it provides an invaluable source of insight into the effect of the formal, conventional nature of language on efforts to think about the nature of the boundary between humans and machines. There is yet another metaphor to describe the traditional research program in AI: "The computer is the physical embodiment of the symbolic calculations envisaged by Hobbes and Leibniz. As such, it is really not a thinking machine but a *language machine*" (Winograd 1991: 216). When the AI project is understood in this way, the computer-as-mind metaphor points to the level of information-processing and symbolic manipulation, not to the more general concept of "thinking."

The metaphor of the computer as a language machine makes sense of the boundary metaphor by locating the boundary more accurately, within the realm of "verbal agreement." One can still ask whether this claim does not also beg the question of material differences in the manner of the Turing test, however. There is more to natural language than the processing of symbols, more than conventional "rules and propositions" that lead to "verbal agreement." If the notion of "symbol system" is indeed "inherently linguistic" (ibid.), everyday natural human language, on the other hand, cannot simply be reduced to the conventional manipulation of symbols.

Hubert L. Dreyfus has stated this objection regularly since 1972: there are things that computers (still) can't do because they function in a binary logic at odds with human reasoning, and binary translations into machine logic of symbols are far from enough to mimic human thinking. AI has been at the same time overly ambitious in its claim to model human intelligence and insufficiently ambitious in trying to understand the linguistic phenomenon and the path it opens to the body. Engelbart, following Whorf, however, was able to see the ways in which the analogous character of natural language, thought, and the human body meant that as a "language machine," the computer could serve as a genuine boundary-spanning object. In this perspective, the materiality of humans and computers takes on a different meaning than that of a "false residual" in a metaphor: both language and technology

are inherently tied to the body on the human side of the interface and to circuits on the electronic side.

The Computer as a Medium and the Question of the Interface

The introduction of visual metaphors has certainly been the most important aspect of the efforts to design adequate user interfaces, understood here as adequate patterns of interaction between the user and the computer. The opening of the graphic dimension of the computer as a communication medium is often thought to be one of the major contributions of Alan Kay and his team at Xerox Palo Alto Research Center (PARC) in the 1970s. A major contribution of this outstanding set of computer scientists is the "desktop metaphor," the leading metaphor for the personal computing interface, as a visual space populated with iconic representations and organized in "windows." Let us see now how the notion of the hypertextual interface unfolds in more specific models of the features of this symbolic space.

In her remarkable *Computer as Theater*, Brenda Laurel (1991: 12–14), narrates how she and the participants of a seminar at the Atari Company (where she was then working) attempted to define the user interface. They rapidly dismissed the simplest model of the interface represented as the space between the person (user) and the computer, that "encompasses what appears on the screen, hardware input/output devices, and their drivers." They dismissed "this over-simplistic model" because of its lack of consideration for the "person's 'mental model' of the computer and the computer's 'understanding' of the person [which] are just as much a part of the interface as its physical and sensory manifestations" (ibid.: 12–13).

Once this "precognitive-science" model of the interface is dismissed, the "conceptual interface" is part of the interface. They called the main problem that they encountered then with this updated model of the interface the "horrible recursion": "If you are going to admit that what the two parties 'think' about each other is part of what is going on, you will have to agree that what the two parties think about what the other is thinking about them must perforce be included in the model" (ibid.: 14).

Facing this "nightmare," the seminar turned its attention to "more manageable concepts." They settled on a simpler concept of the user interface: "the interface is that which joins human and computer, conforming to the needs of each." Laurel concluded that this viewpoint "avoids the central issue of what this all means in terms of reality and representation," and that "when we have such trouble defining a concept, it usually means that we are barking up the wrong tree" (ibid.) But it could also merely require a translation of a more complex process of "suspension," or more precisely, of Hegelian *Aufhebung* that "combines the idea of suspension, with its connotation of temporary cessation; transcendence, which suggests a going beyond; and a kind of preservation" (Ashmore 1989: 111). There is even a sign of such a process when Brenda Laurel says that "you can demonstrate Zeno's paradox on the user's side of the barrier until you're blue in the face, but it's only when you traverse it that things get real."

Who is that "you"? It applies to the user as well as to the designer of the interface: the problem is to enable both of them to *act within a representation* (Laurel 1991: 21). The problem in defining the user interface arose in Laurel's narrative with the cognitive-science assumption that representations are part of an interface: the user's representation of the computer, on one hand, and the "computer's understanding of the person" on the other. What do you think the computer understands of the person? Not much in itself, unless "you" anthropomorphize the computer. You understand a lot more, still, if you understand that the computer can represent the designer. In other words, the computer might be able to learn about the user from the representation of the user that the designer of the interface embodies in his/her design.

Following Laurel, we thus realize that the interface is the representational space where user and designer meet, act, and communicate. The anthropomorphization of the computer translates the user's desire for the computer's *responsiveness* and *capacity to perform action*, two major anthropomorphic qualities of the interface, which

"comprise the metaphor of agency" (Laurel 1990: 358). In her defense of anthropomorphism, she stressed later (p. 143) that the quest for these qualities does not lead necessarily to the personification of the computer, but to its invisibility. This invisibility must be the result of a negotiation between user and designer over the competence of the interface agents. The first step in this direction is to realize that human and nonhuman agents in the interface, like characters in a play, cannot be separated from the plot itself.

Characters such as desktop icons or interface agents in general similarly define, and are defined by, the theatrical frame of the interface as a whole. The efficacy of the computer interface depends on developing convincing "characters" in the "narrative" of the user interface. If their negotiation is successful, user and designer reach a consensus on the competence of the agent to perform a task, and the medium disappears in the process: user and designer agree on the "truth" of the representation embodied in the agent, and, consequently, his/hers/its action appears as "real." The object of the negotiation is the plot, and in the present case, the alternative representations of the user and the designer of the task to be performed with the help of the computer.

Being the result of a consensus between designer and user, the interface agent combines the two orthogonal dimensions of a representation, delegation and inscription. *Delegation* is the process by which the agent is granted the right to represent action in the interface, and *inscription* is the process that enables the agent to perform the action being represented. These two processes conjointly define the competence of the agent as the embodiment of the dialogical consensus reached by users and designers. Michel Serres enables us to shade a new light on this process when he states that "*To hold a dialogue is to suppose a third man and to seek to exclude him*" (1982: 66–7, emphasis in original). Following Serres it is not the medium that disappears in the successful negotiation of the competence of the interface agents, but the "third man," the source of noise. In the case of a successful negotiation between users and designers, the interface agents are under control, they will not create noise but docilely perform what they are expected to.

In this sense, the degree of interactivity of the interface can be seen as the relative opportunity for both user and designer to take part in the two dimensions of the representation process. The joint construction of the plot is the consensus reached on a set of agents whose competences are negotiated between user and designer. Setting the negotiation at the level of the entire set of agents allows one to focus on the representativeness of the interface as a whole, that is, as a socially constructed narrative between user and designer.

The Designer as the Third Man

There were three main shifts in our (occidental) conception of the dialogue. The origin-point is of course the Platonic dialogue. The first shift occurred at the dawn of experimental science, and has been beautifully described and explained by Shapin and Shaffer (1985) as the "virtual witnessing." Dialogue then occurs between the scientist and the world, and is inscribed for the readers in the text: it is therefore, (1) *a dialogue between the scientist (author) and the world (nature) by means of experiments, virtually witnessed by the readers (other scientists, peers).*

The second shift in our conception of dialogue occurred with Cybernetics, the science of communication and control, and more precisely with its theory of information. An alternative dialogue emerged, a dialogue where, according to Serres (1982: 66), authors and readers could trade places, and unite against the noise, "the set of these phenomena of interference that become obstacles to communication." In this third conception of the dialogue, the first candidate for exclusion was the World: "in order for dialogue to be possible, one must close one's eyes and cover one's ears to the song and the beauty of the sirens" (ibid.: 70). The dialogue hence becomes (2) *a dialogue between the author and his or her readers, excluding the world, that is, the object of his or her representative practice.*

Here comes up against reflexive problem: when (1) is deconstructed by means of (2) one faces the specter of the infinite regress (Ashmore 1989), since in principle there could always be a

possible deconstruction of (2) itself by the same process that deconstructed (1) in the first place. The trick is here, of course, all included in "∞," the sign standing for the "infinite" and/or its alternative expression "in principle . . . always . . . possible." In practice now, this iteration of the same basic operation *takes time*. Or, as Rotman (1993: 147, emphasis in original) puts it, "the response to 'infinity' developed here can be summarized as hinging on the issue of *embodiment*." If there is a dialogue then, it takes place between the readers and the World. The limit of the embodiment of the author is the first interface in the set-up that unites author and reader in the same fiction, because and against the virtuality of the physics (not the metaphysics) of their ever-allusive co-presence. The third shift in our conception of dialogue, maybe yet to come, would lead us then to (3) *a dialogue between the readers and the World, partly witnessed and then wholly imagined by the author, in the practice that Malcolm Ashmore called "wrighting."* For Ashmore, "celebratory practical reflexive inquiry is wrighting beyond the *tu quoque*." This celebration, indeed, "is not a matter of authorial presentation . . . not a matter of being correct . . . not a matter of meta-analysis . . . not a matter of solving a problem" (ibid.: 110).

This scheme can be applied to the idea of a dialogue between designer (author) and users (readers) mediated by the hyper(textual) interface. Each of the three conceptions of the dialogue previously introduced corresponds to a specific way to characterize the interaction taking place at the interface. The first conception considers the user as an agent engaged in a human–computer interaction, the second conception considers the user as a subject in a computer mediated communicative act against the noise of the world, and the third conception considers the user as a person engaged in his or her own dialogue with the world. In this third conception of the interface as the space of the dialogue between the user and the world, the designer becomes a witness who *imagines* a space for this dialogue, and enables both the world and the user to act within this space: the designer *wrights* the interface.

In this respect, the fundamental question of interface design becomes: "how to wright in order to enable the user to act as a person in his or her dialogue with the world?" Observe that it is through this dialogue that the user engages in the "always unfinished business of human becoming," and therefore that his or her "ability to act as a person" should be considered both as a process and the result of this process. Understood as such, wrighting the interface refers to a *process of personalization*. The designer can thus be conceived as the agent enabling this personalization: his or her action personalizes the interface, that is, it creates the conditions for the user to become a person, to act as such. As we will see, this conception could help us understand a crucial problem in the evolution of the user interface from today's direct manipulation interfaces to the possible intelligent user interfaces of tomorrow, and, of more concern to our topic, to the hypertextual interface.

The Distribution of Intelligence at the Interface and the Future of the Person

An important debate in today's negotiations about the future of the computer still concerns the question of the distribution of intelligence at the interface, what most analyses have translated into the question of the *personification* of the computer. But today's personification of the computer still requires a "momentary suspension of disbelief" (Laurel 1991: 113). Why should the users temporarily suspend their disbelief? In the case of theater or fiction, Coleridge argued that it is for the sake of experiencing other emotional responses, in Aristotelian terms *catharsis*, the pleasurable release of emotion. After Coleridge, Laurel (ibid.: 120–2) argues that the same process should occur in interacting with a computer, where catharsis stems from achieving a given task.

When a play fails, when the spectators do not feel the release of emotion associated with the characters' experiences, they usually blame the director or the actors. When the interface fails, most of the users of current personal computers blame themselves. This might be the clearest marketing success of the personal computing industry, but it is still an obstacle to the diffusion of its products. Proponents of the personification

of the computer, however, hold that it should not be the case. In the best tradition of the Artificial Intelligence program, they ask the users to grant a momentary suspension of disbelief to the interface agent in the narrative space of the interface, and, at the same time, to consider the agent as a person enabling them to carry out a task in the "real, situated space" of their work practices. Both spaces merge in the "personal space" of the user, and this merger creates a fundamental problem for the request to suspend disbelief: it is the source of breakdowns.

This fundamental problem sometimes leads the proponents of the personification of the computer to make ambiguous claims. On the one hand, the personification of the computer, via the interface agent, is supposed to lead to its disappearance: the medium is supposed to vanish, be "ready at hand," and be an extension of the body. On the other hand, the personification of the agent makes it a part of the "world of people and information" and creates identification for the user: in this second sense, the *medium* becomes "present at hand." It draws attention to itself. Hence, the interface agent is necessarily a source of breakdowns, of never-ending shifts of the attention of the user. Thus, the personification of the interface does not solve the problem of breakdowns; rather, it welds an essential source of breakdowns into the interface. The necessity of the user's "suspension of disbelief" becomes reiterated with each breakdown. Because of its hybrid nature as a symbolic and material space, the user interface is a space of breakdowns, and the personification of the computer cannot prevent such breakdowns by design: it still requires learning and adaptation on the part of the user. Let us see now what an alternative notion of "personalization" could bring to this question.

So far, we have mostly limited ourselves to a narrative conception of the interface. We have seen the drawbacks of current alternatives to reshape the interface, and insisted on the limits of a metaphorical conception. It is now time to complete our review of the question of the future of the interface by insisting on a phenomenological perspective on the "person."

Like Jaron Lanier (1995), "I have long believed that the most important question about information technology is 'How does it effect our definition of what a person is?' " According to Alfred North Whitehead, the process of personalization is the temporal and *continuous* process which constitutes the unity of the subject. The continuous character of the process must be emphasized because for Whitehead, "a nexus enjoys 'personal order' when (a) it is a 'society,' and (b) when the genetic relatedness of its members orders these members 'serially' " (1929: 34). This "genetic relatedness" is the engine of the process.

Whitehead's conceptions had a profound impact on Gregory Bateson's conceptualization of an "ecology of mind," his framework for the exploration of the "natural history of the relationships between explicit, implicit and embodied ideas in the world of living things." Whitehead's influence is especially apparent in Bateson's notion of "socialization," a keystone in his early work in anthropology. Bateson holds that "socialization (by definition) requires *interaction*, usually of two or more organisms" and that its goal is "supposedly" this "set of appearances" that we call a "person," since a "person" is necessarily a socialized individual (Bateson 1975: 75). Bateson made this point clearer in a footnote where he notes that "the 'person' after all, is the *mask*. It is what is perceivable of a human organism. It is a unilateral view of the interface between one organism and another" (ibid.).

In this respect, Bateson's notion of the "person" actually dates back to the original Latin notion of *persona*, in the sense of the mask of tragedy, whose meaning was reconstructed by the Latin etymologists as *pers/sonare*, the mask through (*per*) which sounds (*sonare*) the actor's voice (Mauss, [1938]: 350); but it is also the mask of wax cast on the face of the dead ancestor (*imago*), the *prosopon* of the Greek tradition, which also meant the mask of the ancestor or his statue kept in the wings of the family house (ibid.: 352). Marcel Mauss thus insisted on the original double meaning of the *persona*, that both hides and reveals the true nature of the individual, in respect to his origins, his genetic relatedness to his ancestors, and his singularity as an actual entity.

Hence, personalization and socialization are the two sides of the same process of human becoming. Personalization puts the emphasis on

the appearance of the unity of the subject, while socialization stresses the relatedness of members in their expression of a common form. The are both basically grounded in two dynamics: (1) a continuous dynamic of genetic relatedness stemming from an origin, and (2) a dynamic of feeling which constitutes a common sense for this origin.

Bertrand Russell, Whitehead's former pupil and collaborator, held that there were two ways to define a person: (1) in derivation of memory, since "each person's experience is private to himself, and when one experience consists of recollecting another, the two are said to belong to the same 'person'"; and (2) in derivation of the body, since "we can then define a 'person' as the series of mental occurrences connected with a given body" (Russell 1935: 140). These two derivations are the two projections which anchor Rotman's tripartite scheme: the Agent is active is the field of memory, the Subject is active in a field of mental occurrences connected to a given body, and the Person transcends and unites both in the field of "consciousness." "Consciousness" is an afterthought: it is the result of an articulation of the two previous elementary processes, experience and order, it is the name human beings give to the realization of their dual nature as persons. Now, how on earth could hypertext affect this process of human becoming?

Hypertext, Cybernetics, and Space-time

In an essay written in 1939 and entitled "The Relation of Habitual Thought and Behavior to Language," Benjamin Lee Whorf introduces his inquiry as follows:

> That portion of the whole investigation here to be reported may be summed up in two questions: (1) Are our own concepts of time, space, and matter given in substantially the same form by experience to all men, or are they in part conditioned by the structure of particular languages? (2) Are there traceable affinities between (a) cultural and behavioral norms and (b) large-scale linguistic patterns? (Whorf 1941)

The answers Whorf gave to these questions are considerably more nuanced than the bold formulations, such as "language is culture," that sometimes are attributed to him: "I should be the last to pretend that there is anything so definite as a 'correlation' between culture and language . . . We have plenty of evidence that this is not the case" (ibid.). Whorf's answer to the question of whether our concepts of time, space, and matter are universal and unconditioned or "in part conditioned by the structure of particular languages" was that both propositions are true: space may indeed be perceived in a similar fashion by every individual, and therefore be common to all human beings as a result of the basic conditions of human physiology; while at the same time the concept of space is also a linguistic construction and therefore varies with the different human groups singularized by their language.

Thus, for Whorf, the connections between language, cultural norms, and behavior are to be found at the level of observation and representation, not the level of perception. However unconscious the part that language plays in this process may be, Whorf postulated that it always plays a central role in constituting the "real world" through the process of sharing meaning: "Concepts of time and matter are not given in substantially the same form by experience to all men but depend upon the nature of language or languages through the use of which they have been developed" (ibid.: 159).

The relativity of time and space, of course, was not a notion limited to Whorf's work on comparative linguistics. It was fundamental to cybernetics too (see Chapter 14, Cybernetics). Norbert Wiener likewise insisted on alternative conceptions of time, although not exactly in the perspective presented here by Whorf. He wrote (1948: 37) that "it is thus not too much to say that not only the Newtonian astronomy but even the Newtonian physics has become a picture of the average result of a statistical situation, and hence an account of an evolutionary process." This point is crucial to cybernetics because it eventually justified the functional analogy between living organisms and machines, and hence between brains and computers. Wiener reinforced this idea when he stated that "the great contribution of Heisenberg to physics was the replacement

of this still quasi-Newtonian world of Gibbs by one in which time series can in no way be reduced to an assembly of determinate threads of development in time" (ibid.: 92).

The relativist revolution in how time and space are conceived helped make possible insights such as Whorf's, most notably via the "general semantics" program epitomized by Alfred Korzybski's writings. Korzybski established a connection between the changing worldview in physics and an overall framework for social sciences that granted a new epistemological status to meaning and language. In fact, Whorf's "Relation of Habitual Thought and Behavior to Language" actually appeared in *ETC: A Review of General Semantics*, the General Semantics International Society's journal. For Korzybski (1926), "mankind" is not bound by time, but instead a "*time-binding*" class of life that has survived in evolution by its ability to learn from past experiences and to pass this knowledge on from generation to generation through language. For Korzybski and his followers, it was necessary to change the linguistic conception of the relationship between the "word" and the "thing-in-the-world." Korzybski (1933) best expressed this new perspective in his famous analogy between maps and language: "a map *is not* the territory" to which it corresponds, "words are not the things they represent." This central premise led him to question the fundamental basis of the Newtonian worldview and the Aristotelian system of logic, and to propose, instead, a "relativist" reformulation of the law of identity: "*identity is a relative matter*: relative to the history of the things considered, relative to the environment the thing is in, relative to our own practical purposes, relative to the frame of reference from which it is viewed" (Reiser 1989: 85–6).

As in the shifting worldview in physics, the extension of the Aristotelian logic and its application to language in the General Semantics program relied heavily on a quasi-statistical approach, or more accurately, on a theory of classes, since words were not considered any more as "identical" with what they represented, but rather as a class of things that they *could* represent. The social scientists of the cybernetics group were of course aware of this line of thought. The connection between Korsybski's ideas and the social science side of cybernetics appears in the synthesis provided by Gregory Bateson's writings. The core of this synthesis revisited Korzybski's notion of the connection between "map" and "territory," and applied it at a basic level:

> The bridge between map and territory is *difference*. It is only *news of difference* that can get from the territory to the map, and this fact is the basic epistemological statement about the relationship between all reality out there and all perception in here: that the bridge must always be in the form of difference. (Bateson 1979: 240)

Bateson also redefined the basic cybernetic notion of "information" as "any difference that makes a difference" (1979: 228). Defined in these terms, for Whorf, what matters is not simply that language determines our concepts and that the relational network of our concepts (links, connections) determines our view of the world. What matters is the difference between perception on the physiological level and mental concepts on the level of language, a difference that itself has different implications, depending on which side of it is emphasized. For Korzybski, what matters is the difference between the map and the territory – not the map by itself, or the territory. Difference is thus by definition the site of an interface. In the formulations of Whorf, Korzybski, and Bateson, the interface is this difference that makes a difference, at the boundary between the physical and the conceptual realms.

Conclusion

The computer will be "personal" when it allows its users to act as persons, to experience the world as both embodied subject and ideal agent, united in actual socialized and enduring persons. This could happen if and only if the computer could affect, in terms of signification, the way humans express their genetic relatedness through their "mutual prehensions" of each other. These prehensions are the inscribing and incorporating practices which characterize the human experience in the fields of virtual *and* actual entities.

However, the computer "memory" is organized quite differently from the human memory: computer "time" is discreet, and corresponds to an event-oriented worldview, anchored in the renewed notion of probabilistic time brought in by the relativist tradition. This worldview seems at first to contradict the requisite of continuity. Computer events are addresses without a fixed seriality: there are the discreet traces of previous interactions, and because of that they refer only to virtual entities. The fact that a human being could imagine a continuous temporality in his hermeneutical relation to the computer does not change the fact that the computer does not share it as an embodiment relation: this continuity is only an interpretation of the human user and of his interaction with the computer.

Could this interpretation be embodied in the computer in the same the way that it is in human beings? A starting-point would be the ability of the computer to allow humans to make full use of their body, to act inside the space that the interface presents *and* represents, to feel and experience tomorrow's worlds as bodies and minds, through the symbolic and material space that an interface affords for exploration, play, and work. The social and cultural construction of the personal computer user, so far, has led to an overwhelming hegemony of the visual sense and symbolic coding, following in that a larger trend in modernity. Whether the computer could be interpreted as a person or as slave, whether it could be metaphorically considered as a conscious entity or not, is finally of little interest in regard to what the computer *does*. Nothing prevents *a priori* the computer from participating in the ongoing evolution of human beings, in the way that human tools have done for a long time. Conscious efforts to design personal computers should take this into account and strive for a harmonious human experience, in its fullest expression. So far, the personal interface has remained a marking interface. Typing and clicking is marking, indexing, punching a hole. But this despairing notion might hide the great potential of the hypertextual interface: a system that could allow human action in cybernetic space-time, through the linked events of human becoming, in language and gestures, in thought and reality.

Acknowledgments

This chapter is a revised version of a paper entitled "Bridging the gulfs: from hypertext to cyberspace" that appeared in the *Journal of Computer-mediated Communication* in september 1997 (<http://207.201.161.120/jcmc/vol3/issue2/coverpage.shtml>). Elements of the final sections also appeared in *Boot-strapping: Douglas Engelbart, Coevolution and the Origins of Personal Computing* (2000). I thank Luciano Floridi for his interest in these previous works and for his help in editing the present chapter.

References

Ashmore, M. 1989. *The Reflexive Thesis: Wrighting Sociology of Scientific Knowledge*. Chicago: University of Chicago Press. [Advanced undergraduates, graduates.]

Bateson, G. 1979. *Mind and Nature: A Necessary Unity*. New York: Dutton. [Advanced undergraduates, graduates.]

——. 1991 [1975]. "Some components of socialization for trance." In R. E. Donaldson, ed., *Sacred Unity: Further Steps to an Ecology of Mind*. New York: HarperCollins, pp. 73–88. [Advanced undergraduates, graduates.]

Bush, V. 1991 [1945]. "As we may think." In J. M. Nyce and P. Kahn, eds., *From Memex to Hypertext: Vannevar Bush and the Mind's Machine*. San Diego, CA: Academic Press, pp. 83–107. (Originally published in *Atlantic Monthly* 176[1]: 641–9.) [Undergraduates, graduates.]

Engelbart, D. C. 1962. "Augmenting human intellect: a conceptual framework." Report to the Director of Information Sciences, Air Force Office of Scientific Research, Menlo Park: Stanford Research Institute, Oct. [Advanced undergraduates, graduates.]

Floridi, L. 1999. *Philosophy and Computing: An Introduction*. London: Routledge. [Undergraduates, graduates.]

Gygi, K. 1990. "Recognizing the symptoms of hypertext . . . and what to do about it." In B. Laurel, ed., *The Art of Human Computer Interface Design*. Reading, MA: Addison-Wesley, pp. 279–87. [Undergraduates, graduates.]

Korzybski, A. 1926. *Time-binding: The General Theory*. Washington, DC: J. C. Wood. [Graduates.]

——. 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics.* Lancaster, PA: Science Press. [Advanced undergraduates, graduates.]

Landow, G. P. 1992. *Hypertext: The Convergence of Contemporary Critical Theory and Technology.* 2nd ed. 1997. Baltimore: Johns Hopkins University Press. [Undergraduates, graduates.]

Lanier, J. 1995. "Agents of alienation." *Journal of Consciousness Studies* 2: 76–81. [Undergraduates, graduates.]

Laurel, B. 1990. "Interface agents: metaphors with character." In B. Laurel, ed., *The Art of Human Computer Interface Design.* Reading, MA: Addison-Wesley, pp. 355–65. [Undergraduates, graduates.]

——. 1991. *Computers as Theater.* Reading, MA: Addison-Wesley. [Undergraduates, graduates.]

Mauss, M. 1950. [1938] "Une catégorie de l'esprit humain: la notion de personne celle de 'moi.'" In *Sociologie et anthropologie.* Paris: Presses Universitaires de France, pp. 331–62. [Advanced undergraduates, graduates.]

Newell, A. 1991. "Metaphors for mind, theories of mind: should the humanities mind?" In J. J. Sheehan and M. Sosna, eds., *The Boundaries of Humanity: Humans, Animals, Machines.* Berkeley, CA: University of California Press, pp. 158–97. [Advanced undergraduates, graduates.]

Nyce, J. M. and Kahn, P. 1991. "A machine for the mind: Vannevar Bush's memex." In J. M. Nyce and P. Kahn, eds., *From Memex to Hypertext: Vannevar Bush and the Mind's Machine.* San Diego, CA: Academic Press, pp. 39–66. [Undergraduates, graduates.]

Reiser, O. L. 1989. "Historical-cultural significance of non-Aristotelian movement and the methodological contribution of Korzybski." In S. I. Berman, ed., *Logic and General Semantics: Writings of Oliver L. Reiser and Others.* San Francisco: International Society for General Semantics, pp. 89–97. [Advanced undergraduates, graduates.]

Rotman, B. 1993. *Ad Infinitum: The Ghost in the Turing Machine: Taking God out of Mathematics and Putting the Body Back In.* Stanford: Stanford University Press. [Graduates.]

Russell, B. 1997 [1935]. *Religion and Science.* Oxford: Oxford University Press. [Undergraduates, graduates.]

Serres, M. 1982. "Platonic dialogue." In J. V. Harari and D. F. Bell, eds., *Hermes: Literature, Science, Philosophy.* Baltimore, MD: Johns Hopkins University Press, pp. 65–70. [Advanced undergraduates, graduates.]

Shapin, S. and Schaffer, S. 1985. *Leviathan and the Air-pump: Hobbes, Boyle, and the Experimental Life.* Princeton, NJ: Princeton University Press. [Advanced undergraduates, graduates.]

Whitehead, A. N. 1978 [1929]. *Process and Reality.* New York: Free Press. [Graduates.]

Whorf, B. L. 1956 [1927] "On the connection of ideas." In J. B. Carroll, ed., *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* Cambridge, MA: MIT Press, pp. 35–9. [Undergraduates, graduates.]

——. 1956 [1941] "The relation of habitual thought and behavior to language." In J. B. Carroll, ed., *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* Cambridge, MA: MIT Press, pp. 134–59. [Advanced undergraduates, graduates.]

Wiener, N. 1961 [1948]. *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, MA: MIT Press. [Advanced undergraduates, graduates.]

Winograd, T. 1991. "Thinking machines: can there be? Are we?" In J. J. Sheehan and M. Sosna, eds., *The Boundaries of Humanity: Humans, Animals, Machines.* Berkeley, CA: University of California Press, pp. 198–223. [Advanced undergraduates, graduates.]

Part VI

Logic and Probability

Chapter 20

Logic

G. Aldo Antonelli

Origins of the Modern Conception of Logic

Logic is an ancient discipline that, ever since its inception some 2500 years ago, has been concerned with the analysis of patterns of valid reasoning. Aristotle first developed the theory of the *syllogism* (a valid argument form involving predicates and quantifiers), and later the Stoics singled out patterns of *propositional* argumentation (involving sentential connectives). The study of logic flourished in ancient times and during the middle ages, when logic was regarded, together with grammar and rhetoric (the other two disciplines of the *trivium*), as the foundation of humanistic education.

Throughout its history, logic has always had a *prescriptive* as well as a *descriptive* component. As a descriptive discipline, logic aims to capture the arguments accepted as valid in everyday linguistic practice. But this aspect, although present throughout the history of the field, has since the inception of the modern conception of logic, some 100 or 150 years ago, taken up a position more in the background, and in fact some have argued that it is no longer part of logic proper, but belongs to other disciplines (linguistics, psychology, or what have you). Nowadays logic is, first and foremost, a prescriptive discipline, concerned with the identification and justification of valid inference forms.

The articulation of logic as a prescriptive discipline is, ideally, a two-fold task. The first task requires the identification of a class of valid arguments. The class thus identified must have certain features: not just any class of arguments will do. For instance, it is reasonable to require that the class of valid argument be closed under the relation "having the same logical form as," in that if an argument is classified as valid, then so is any other argument of the same logical form. It is clear, then, that such an identification presupposes, and rests on, a notion of *logical form*.

The question of what constitutes a good theory of logical form exceeds the boundary of the present contribution, and hence we will not be concerned with it. We shall limit ourselves to the observation that one can achieve the desired closure conditions by requiring that the class of valid arguments be generated in some uniform way from some restricted set of principles. For instance, Aristotle's theory of the syllogism accomplishes this in a characteristically elegant fashion: subject–predicate propositions are classified on the basis of their forms into a small number of classes, and syllogisms are then generated by allowing the two premises and the conclusion to take all possible forms.

The second task, however, is much harder. Once a class of arguments is identified, one naturally wants to know what it is that makes these arguments *valid*. In other words, in order to accomplish this second task, one needs a general theory of *logical consequence*, and such a theory was not only unavailable to the ancients, it would not be available until the appearance of modern symbolic logic in the late 1800s, when an effort was undertaken to formalize and represent mathematical reasoning, and it would not be completely developed until the middle of the twentieth century.

It is only with the development of the first general accounts of the notion of logical consequence that modern symbolic logic was born. Modern symbolic logic is only a little over 100 years old, dating back to the end of the nineteenth century, and in particular to Gottlob Frege's *Begriffsschrift* (1879) and Richard Dedekind's *Was sind und was sollen die Zahlen?* (1888). With these two works we have the beginning of a rigorous account of logical consequence, an account that will be perfected by Alfred Tarski in the early 1930s.

This chapter focuses on the development of modern symbolic logic from the point of view of the notion of *logical consequence*. After presenting a streamlined account of what is regarded as the crowning achievement of modern symbolic logic, i.e., the systematization of *first-order logic*, we consider consequence relations from an abstract point of view. In the next section we look at consequence relations that are of particular conceptual interest, in that they aim to capture patterns of *defeasible* reasoning in which conclusions are drawn tentatively, subject to being retracted in the light of additional evidence. Finally, we look at nonmonotonic logics devised to capture such defeasible inference.

First-order Logic

First-order logic (henceforth: FOL) was originally developed (through the work of Frege, Dedekind, Russell & Whitehead, Hilbert, Gödel, and Tarski) for the representation of mathematical reasoning. As such, FOL turned out to be nothing but a stunning success. Its mathematical properties provide a crucial benchmark for the assessment of alternative logical frameworks. We are not going, in this chapter, to provide an introduction to the nuts and bolts of FOL: the interested reader can consult any one of the many excellent introductory texts that are available, such as, for example, Enderton 1972.

FOL provides an implementation of the so-called "no-counterexample" consequence relation: a sentence ϕ is a consequence of a set Γ of sentences if and only if one cannot reintepret the language in which Γ and ϕ are formulated in such a way as to make all sentences in Γ true and ϕ false. An inference from premises $\psi_1, \ldots, \psi_k$ to a conclusion ϕ is *valid* if ϕ is a consequence of $\{\psi_1, \ldots, \psi_k\}$, i.e., if the inference has no counterexample.

For this to be a rigorous account of logical consequence, the underlying notion of interpretation needs to be made precise. This was accomplished by Alfred Tarski in 1935, who defined the notion of truth on an interpretation (see Tarski 1956 for a collection of his technical papers). In so doing, Tarski overcame both a technical and a philosophical problem. The technical problem has to do with the fact that in FOL quantified sentences are obtained from components that are not, in turn, sentences, so that a direct recursive definition of truth for sentences breaks down at the quantifier case. In order to overcome this problem Tarski introduced the auxiliary notion of *satisfaction*. The philosophical obstacle had to do with the fact that the notion of *truth* was considered suspiciously metaphysical among logicians trained in the environment of the Vienna Circle. This was a factor, for instance, in Gödel's reluctance to formulate his famous undecidability results in terms of truth.

Tarski's analysis yielded a mathematically precise definition for the "no-counterexample" consequence relation $\models$ of FOL: we say that ϕ is a consequence of a set Γ of sentences, written $\Gamma \models \phi$, if and only if ϕ is true on every interpretation on which every sentence in Γ is true. At first glance, there would appear to be something intrinsically infinitary about $\models$. Regardless of whether Γ is finite or infinite, to check whether $\Gamma \models \phi$ one has to "survey" infinitely many possible interpretations, and check whether any one

of them is a counterexample to the entailment claim, i.e., whether any one of them is such that all sentences in Γ are true on it while ϕ is false.

However, surprisingly, in FOL the infinitary nature of $\models$ is only apparent. As Gödel showed in 1929, the relation $\models$, although defined by universally quantifying over all possible interpretations, can be analyzed in terms of the existence of finite objects of a certain kind, viz., formal proofs. A *formal proof* is a finite sequence of sentences, each one of which is either an *axiom*, or an *assumption*, or is obtained from previous ones by means of one of a finite number of inference rules, such as *modus ponens*. Many different axiomatizations of FOL exist, and a particularly simple and elegant one can be found in Enderton 1972. If a sentence ϕ occurs as the last line of a proof, then we say that the proof is a *proof of* ϕ; and we say that ϕ is *provable from* Γ, written $\Gamma \vdash \phi$, if and only if there is a proof of ϕ all of whose assumptions are drawn from Γ.

Gödel's famous completeness theorem of 1929 states that the two relations $\models$ and $\vdash$ are extensionally equivalent: for any ϕ and Γ, $\Gamma \models \phi$ if and only if $\Gamma \vdash \phi$. This is a remarkable feature of FOL, which has a number of consequences. One of the deepest consequences follows from the fact that proofs are finite objects, and hence that $\Gamma \vdash \phi$ if and only if there is a *finite* subset Γ_0 of Γ such that $\Gamma_0 \vdash \phi$. This, together with the completeness theorem, gives us the *compactness theorem*: $\Gamma \models \phi$ if and only if there is a finite subset Γ_0 of Γ such that $\Gamma_0 \models \phi$. There are many interesting equivalent formulations of the theorem, but the following is perhaps the most often cited. Say that a set of sentences is *consistent* if they can all be made simultaneously true on some interpretation; then the compactness theorem says that a set Γ is consistent if and only if each of its finite subsets is by itself consistent.

Another important consequence of Gödel's completeness theorem is the following form of the Löwenheim–Skolem theorem: if all the sentences in Γ can be made simultaneously true in some interpretation, then they can also be made simultaneously true in some (other) interpretation whose universe is no larger than the set $\mathbb{N}$ of the natural numbers.

Together, the compactness and the Löwenheim–Skolem theorem are the beginning of one of the most successful branches of modern symbolic logic: model theory. Compactness and the Löwenheim–Skolem theorem characterize FOL; as shown by Per Lindström in 1969, any logical system (meeting certain "regularity" conditions) for which both compactness and Löwenheim–Skolem hold, is no more expressive than FOL (see Ebbinghaus et al. 1994: ch. 13 for an accessible treatment).

Another important consequence of Gödel's completeness theorem has to do with the question of whether and to what extent one can devise an effective procedure to determine if a sentence ϕ is valid, or, more generally, whether $\Gamma \models \phi$ for given Γ and ϕ. First, some terminology. We say that a set Γ of sentences is *decidable* if there is an effective procedure, i.e., a mechanically executable set of instructions, that determines, for each sentence ϕ, whether ϕ belongs to Γ or not. Notice that such a procedure gives both a positive and a negative test for membership of a sentence ϕ in Γ. A set of sentences is *semidecidable* if there is an effective procedure that determines if a sentence ϕ is a member of Γ, but might not provide an answer if ϕ is not a member of Γ. In other words, Γ is semidecidable if there is a positive, but not necessarily a negative test for membership in Γ. Equivalently, Γ is semidecidable if it can be given an effective listing, i.e., if it can be mechanically generated. These notions can be generalized to relations among sentences of any number of arguments. For instance, it is an important feature of the axiomatizations of FOL, such as that of Enderton 1972, that both the set of axioms and the relation that holds among $\phi_1, \ldots, \phi_k$ and ψ when ψ can be inferred from $\phi_1, \ldots, \phi_k$ by one of the rules, are decidable. As a result, the relation that holds among $\phi_1, \ldots, \phi_k$ and ϕ whenever $\phi_1, \ldots, \phi_k$ is a proof of ϕ is also decidable. (See Chapter 2, COMPUTATION, for further details on these notions.)

The import of Gödel's completeness theorem is that if the set Γ is decidable (or even only semidecidable), then the set of all sentences ϕ such that $\Gamma \models \phi$ is semidecidable. Indeed, one can obtain an effective listing for such a set by systematically generating all proofs from Γ. The question arises of whether, beside this positive test, there might not also be a negative test for a

sentence ϕ being a consequence of Γ. This "decision problem" (*Entscheidungsproblem*) was originally proposed by David Hilbert in 1900, and it was solved in 1936 independently by Alonzo Church and Alan Turing. The Church–Turing theorem states that, in general, it is not decidable whether $\Gamma \models \phi$, or even if ϕ is valid. (It is important to know that for many, even quite expressive, fragments of first-order logic the decision problem is solvable; see Börger et al. 1997 for details.) We should also notice the following fact that will be relevant later in our discussion: say that a sentence ϕ is *consistent* if $\{\phi\}$ is consistent, i.e., if its negation $\neg\phi$ is not valid. Then the set of all sentences ϕ such that ϕ is consistent is not even semidecidable, for a positive test for such a set would yield a negative test for the set of all valid sentences, which would thus be decidable, against the Church–Turing theorem.

Consequence relations

In the previous section, we defined a consequence relation $\models$ by saying that $\Gamma \models \phi$ if and only if ϕ is true on every interpretation on which every sentence in Γ is true. In general, it is possible to consider what abstract properties a relation of consequence between sets of sentences and single sentences could have. Let $\vdash$ be any such relation. Consider the following properties, all of which are satisfied by the consequence relation $\models$ of FOL:

Supraclassicality: if $\Gamma \models \phi$ then $\Gamma \vdash \phi$.
Reflexivity: if $\phi \in \Gamma$ then $\Gamma \vdash \phi$;
Cut: If $\Gamma \vdash \phi$ and $\Gamma, \phi \vdash \psi$ then $\Gamma \vdash \psi$;
Monotony: If $\Gamma \vdash \phi$ and $\Gamma \subseteq \Delta$ then $\Delta \vdash \phi$.

The first property is supraclassicality, which states that if ϕ follows from Γ in FOL, then it also follows according to $\vdash$; i.e., $\vdash$ extends $\models$ (the relation $\models$ is trivially supraclassical). Of the remaining conditions, the most straightforward is reflexivity: it says that if ϕ belongs to the set Γ, then ϕ is a consequence of Γ. This is a very minimal requirement on a relation of logical consequence. We certainly would like all sentences in Γ to be inferable from Γ. It's not clear in what sense a relation that fails to satisfy this requirement can be called a *consequence* relation.

Cut, a form of transitivity, is another crucial feature of consequence relations. Cut is a conservativity principle: if ϕ is a consequence of Γ, then ψ is a consequence of Γ together with ϕ only if it is already a consequence of Γ alone. In other words, by adjoining to Γ something which is already a consequence of Γ does not lead to any *increase* in inferential power. Cut is best regarded as the statement that the "length" of a proof does not affect the degree to which the assumptions support the conclusion. Where ϕ is already a consequence of Γ, if ψ can be inferred from Γ together with ϕ, then ψ can also be obtained via a longer "proof" that proceeds indirectly by first inferring ϕ. It is immediate to check that FOL satisfies Cut.

It is worth noting that many forms of probabilistic reasoning fail to satisfy Cut, precisely because the degree to which the premises support the conclusion is inversely correlated to the length of the proof. To see this, we adapt a well-known example. Let Ax abbreviate "x was born in Pennsylvania Dutch country," Bx abbreviate "x is a native speaker of German," and Cx abbreviate "x was born in Germany." Further, let Γ comprise the statements "Most A's are B's," "Most B's are C's," and Ax. Then Γ supports Bx, and Γ together with Bx supports Cx, but Γ by itself does not support Cx. Statements of the form "Most A's are B's" are interpreted probabilistically, as saying that the conditional probability of B given A is, say, greater that 50 percent; likewise, we say that Γ supports a statement ϕ if Γ assigns ϕ a probability $p > 50$ percent.

Since Γ contains "Most A's are B's" and Ax, it supports Bx (in the sense that the probability of Bx is greater that 50 percent); similarly, Γ together with Bx supports Cx; but Γ by itself cannot support Cx. Indeed, the probability of someone who was born in Pennsylvania Dutch country being born in Germany is arbitrarily close to zero. Examples of inductive reasoning such as the one just given cast some doubt on the possibility of coming up with a well-behaved relation of probabilistic consequence (see Chapter 21, Probability in Artificial Intelligence).

Special considerations apply to monotony. Monotony states that if ϕ is a consequence of Γ

then it is also a consequence of any set containing Γ (as a subset). The import of monotony is that one cannot preempt conclusions by adding new premises to the inference. It is clear why FOL satisifies monotony: semantically, if ϕ is true on every interpretation on which all sentences of Γ are true, then ϕ is also true on every interpretation on which all sentences in a larger set Δ are true (similarly, proof-theoretically, if there is a proof of ϕ all of whose assumptions are drawn from Γ, then there is also a proof of ϕ – indeed, the same proof – all of whose assumptions are drawn from Δ).

Many consider this feature of FOL as inadequate to capture a whole class of inferences typical of everyday (as opposed to mathematical or formal) reasoning, and therefore question the descriptive adequacy of FOL, when it comes to representing commonsense inferences. In everyday life, we quite often reach conclusions tentatively, only to retract them in the light of further information. For instance, when told that Stellaluna is a mammal, we infer that she does not fly, because mammals, by and large, don't fly. But the conclusion that Stellaluna doesn't fly can be retracted when we learn that Stellaluna is a bat, because bats are a specific kind of mammals, and they do fly. So we infer that Stellaluna does fly after all. This process can be further iterated. We can learn, for instance, that Stellaluna is a baby bat, and that therefore she does not know how to fly yet. Such complex patterns of *defeasible* reasoning are beyond the reach of FOL, which is, by its very nature, monotonic.

For these and similar reasons, people have striven, over the last 20 years or so, to devise nonmonotonic formalisms capable of representing defeasible inference. We will take a closer look at these formalisms below, but for now we want to consider the issue from a more abstract point of view.

When one gives up monotony in favor of descriptive adequacy, the question arises of what formal properties of the consequence relation to put in its place. Two such properties have been considered in the literature, for an arbitrary consequence relation $\vdash$:

Cautious Monotony: If $\Gamma \vdash \phi$ and $\Gamma \vdash \psi$, then $\Gamma, \phi \vdash \psi$.

Rational Monotony: If $\Gamma \nvdash \neg \phi$ and $\Gamma \vdash \psi$, then $\Gamma, \phi \vdash \psi$.

Both Cautious Monotony and the stronger principle of Rational Monotony are special cases of Monotony, and are therefore not in the foreground as long as we restrict ourselves to the classical consequence relation $\models$ of FOL.

Although superficially similar, these principles are quite different. Cautious Monotony is the converse of Cut: it states that adding a consequence ϕ back into the premise-set Γ does not lead to any *decrease* in inferential power. Cautious Monotony tells us that inference is a cumulative enterprise: we can keep drawing consequences that can in turn be used as additional premises, without affecting the set of conclusions. Together with Cut, Cautious Monotony says that if ϕ is a consequence of Γ then for any proposition ψ, ψ is a consequence of Γ if and only if it is a consequence of Γ together with ϕ. It has been often pointed out by Dov Gabbay that Reflexivity, Cut and Cautious Monotony are critical properties for any well-behaved nonmonotonic consequence relation (see Gabbay et al. 1994, Stalnaker 1994).

The status of Rational Monotony is much more problematic. As we observed, Rational Monotony can be regarded as a strengthening of Cautious Monotony, and like the latter it is a special case of Monotony. However, there are reason to think that Rational Monotony might not be a correct feature of a nonmonotonic consequence relation. A counterexample due to Stalnaker (1994: 19) involves three composers: Verdi, Bizet, and Satie. Suppose that we initially accept (correctly but defeasibly) that Verdi is Italian, while Bizet and Satie are French. Suppose now that we are told by a reliable source of information that Verdi and Bizet are compatriots. This leads us no longer to endorse the propositions that Verdi is Italian (because he could be French), and that Bizet is French (because he could be Italian); but we would still draw the defeasible consequence that Satie is French, since nothing that we have learned conflicts with it. By letting $I(v)$, $F(b)$, and $F(s)$ represent our initial beliefs about the nationality of the three composers, and $C(v, b)$ represent that Verdi and Bizet are compatriots, the situation could be represented as follows:

$C(v, b) \vdash F(s)$.

Now consider the proposition $C(v, s)$ that Verdi and Satie are compatriots. Before learning that $C(v, b)$ we would be inclined to reject the proposition $C(v, s)$ because we endorse and $I(v)$ and $F(s)$, but after learning that Verdi and Bizet are compatriots, we can no longer endorse $I(v)$, and therefore no longer reject $C(v, s)$. The situation then is as follows:

$C(v, b) \nvdash \neg C(v, s)$.

However, if we added $C(v, s)$ to our stock of beliefs, we would lose the inference to $F(s)$: in the context of $C(v, b)$, the proposition $C(v, s)$ is equivalent to the statement that all three composers have the same nationality, and this leads us to suspend our assent to the proposition $F(s)$. In other words, and contrary to Rational Monotony:

$C(v, b), C(v, s) \nvdash F(s)$.

The previous discussion gives a rather clear picture of the desirable features a nonmonotonic consequence relation. Such a relation should satisfy Supraclassicality, Reflexivity, Cut, and Cautious Monotony.

Varieties of Defeasible Reasoning

A separate issue from the formal properties of a nonmonotonic consequence relation, although one that is strictly intertwined with it, is the issue of how *conflicts* between potential defeasible conclusions are to be handled.

There are two different kinds of conflicts that can arise within a given nonmonotonic framework: (i) conflicts between defeasible conclusions and "hard facts"; and (ii) conflicts between one potential defeasible conclusion and another (many formalisms, for instance, provide some form of defeasible inference rules, and such rules might have conflicting conclusions). When a conflict (of either kind) arises, steps have to be taken to preserve or restore consistency.

All defeasible formalisms handle conflicts of the first kind in the same way: indeed, it is the very essence of defeasible reasoning that conclusions can be retracted when new facts are learned. But conflicts of the second kind can be handled in two different ways: one can draw inferences either in a "cautious" or "bold" fashion (also known as "skeptical" or, respectively, "credulous"). These two options correspond to widely different ways to construe a given body of defeasible knowledge, and yield different results as to what defeasible conclusions are warranted on the basis of such a knowledge base.

The difference between these basic attitudes comes to this. In the presence of potentially conflicting defeasible inferences (and in the absence of further considerations such as specificity – see below), the credulous reasoner always commits to as many defeasible conclusions as possible, subject to a consistency requirement, whereas the skeptical reasoner withholds assent from potentially conflicted defeasible conclusions.

A famous example from the literature, the so-called "Nixon diamond," will help make the distinction clear. Suppose our knowledge base contains (defeasible) information to the effect that a given individual, Nixon, is both a Quaker and a Republican. Quakers, by and large, are pacifists, whereas Republicans by and large are not. The question is, what defeasible conclusions are warranted on the basis of this body of knowledge, and in particular whether we should infer that Nixon is a pacifist or that he is not pacifist. Figure 20.1 provides a schematic representation of this state of affairs in the form of a (defeasible) network.

The credulous reasoner has no reason to prefer either conclusion ("Nixon is a pacifist"; "Nixon is not a pacifist") to the other one, but will definitely commit to one or the other. The skeptical reasoner recognizes that this is a conflict not between hard facts and defeasible inferences, but between two different defeasible inferences. Since the two possible inferences in some sense "cancel out," the skeptical reasoner will refrain from drawing either one.

Whereas many of the early formulations of defeasible reasoning have been credulous, skepticism has gradually emerged as a viable alternative, which can, at times, be better behaved. Arguments have been given in favor of both skeptical and credulous inference. Some have

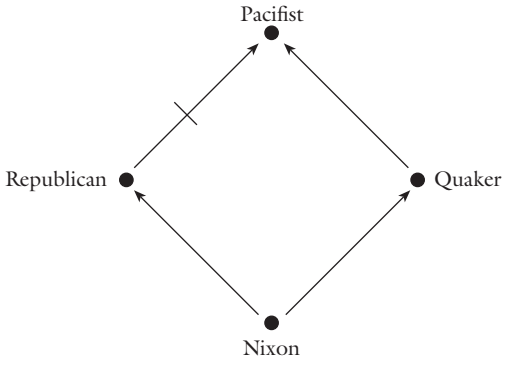

Figure 20.1: The Nixon diamond

argued that credulity seems to better capture a certain class of intuitions, while others have objected that although a certain degree of "jumping to conclusions" is by definition built into any nonmonotonic formalism, such jumping to conclusions needs to be regimented, and that skepticism provides precisely the required regimentation. (A further issue in the skeptical/credulous debate is the question of whether so-called "floating conclusions" should be allowed; see Horty 2002 for a review of the literature and a substantial argument that they should not.)

Nonmonotonic Logics

As we have mentioned, over the last twenty years or so, a number of so-called "nonmonotonic" logical frameworks have emerged, expressly devised for the purpose of representing defeasible reasoning. The development of such frameworks represents one of the most significant developments both in logic and artificial intelligence, and has wide-ranging consequences for our philosophical understanding of argumentation and inference.

Pioneering work in the field of nonmonotonic logics was carried out beginning in the late 1970s by (among others) J. McCarthy, D. McDermott, & J. Doyle, and R. Reiter (see Ginsberg 1987 for a collection of early papers in the field). With these efforts, the realization (which was hardly new) that ordinary first-order logic was inadequate to represent defeasible reasoning was for the first time accompanied by several proposals of formal frameworks within which one could at least begin to talk about defeasible inferences in a precise way, with the long-term goal of providing for defeasible reasoning an account that could at least approximate the degree of success achieved by FOL in the formalization of mathematical reasoning. The publication of a monographic issue of the *Artificial Intelligence Journal* in 1980 can be regarded as the "coming of age" of defeasible formalisms.

The development of nonmonotonic (or defeasible) logics has been guided all along by a rich supply of examples. One of the early sources of motivation for the development of nonmonotonic logic comes from database theory. Consider for instance the *closed world assumption*: suppose that you need to travel from Oshkosh to Minsk, so you consult your travel agent, who, not surprisingly, tells you that there are no direct flights. How does the travel agent know? In a sense, he doesn't: his database does not list any direct flights between Oshkosh and Minsk, and he assumes that the database is *complete*. In other words, what we have in this example is an attempt to *minimize* the extension of a given predicate ("flight-between" in this case). Moreover, such a minimization needs to take place not with respect to what the database explicitly contains but with respect to what it implies.

The idea of minimization is at the basis of one of the earliest nonmonotonic formalisms, McCarthy's *circumscription*. Circumscription makes explicit the intuition that, all other things being equal, extensions of predicates should be *minimal*. Again, consider principles such as "all normal birds fly." Here we are trying to minimize the extension of the abnormality predicate, and assume that a given bird is normal unless we have positive information to the contrary. Formally, this can be represented using second-order logic. In second-order logic, in contrast to FOL, one is allowed to explicitly quantify over predicates, forming sentences such as $\exists P \forall x Px$ ("there is a universal predicate") or $\forall P(Pa \leftrightarrow Pb)$ ("a and b are indiscernible"). In circumscription, given predicates P and Q, we abbreviate $\forall x(Px \rightarrow Qx)$ as $P \leq Q$, and $P \leq Q \wedge Q \not\leq P$ as $P < Q$. If $A(P)$

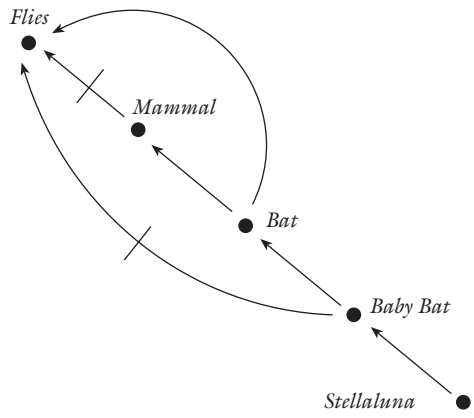

Figure 20.2: An Inheritance network; links of the form $A \to B$ represent the fact that typical A's are B's, and links $A \nrightarrow B$ represent the fact that typical A's are not B's

is a formula containing occurrences of a predicate P, then the circumscription of P in A is the second-order sentence $A^*(P)$:

$$A(P) \wedge \neg\exists Q[A(Q) \wedge Q < P].$$

$A^*(P)$ says that P satisfies A, and that no smaller predicate does. Let Px be the predicate "x is abnormal," and let $A(P)$ be the sentence "All normal birds fly." Then the sentence "Tweety is a bird," together with $A^*(P)$ implies "Tweety flies," for the circumscription axiom forces the extension of P to be empty, so that "Tweety is normal" is automatically true. In terms of consequence relations, circumscription allows us to define, for each predicate P, a nonmonotonic relation $A(P) \vdash \phi$ that holds precisely when $A^*(P) \models \phi$. (This basic form of circumscription has been generalized, for in practice one needs to minimize the extension of a predicate, while allowing the extension of certain other predicates to vary.) From the point of view of applications, however, circumscription has a major shortcoming, namely the absence of a complete inference procedure, due to the fact that, in general, second-order logic lacks such a procedure. The price one pays for the greater expressive power of second-order logic is that there are no complete axiomatizations, as we have for FOL.

Another nonmonotonic formalism inspired by the intuition of minimization of abnormalities is *nonmonotonic inheritance*. Whenever we have a taxonomically organized body of knowledge, we presuppose that subclasses inherit properties from their superclasses: dogs have lungs because they are mammals, and mammals have lungs. However, there can be exceptions, which can interact in complex ways. To use an example already introduced, mammals, by and large, don't fly; since bats are mammals, in the absence of any information to the contrary, we are justified in inferring that bats do not fly. But then we learn that bats are exceptional mammals, in that they do fly: the conclusion that they don't fly is retracted, and the conclusion that they fly is drawn instead. Things can be more complicated still, for in turn, as we have seen, baby bats are exceptional bats, in that they do not fly (does that make them unexceptional mammals?). Here we have potentially conflicting inferences. When we infer that Stellaluna, being a baby bat, does not fly, we are resolving all these potential conflicts based on a *specificity* principle: more specific information overrides more generic information. Nonmonotonic inheritance networks were developed for the purpose of capturing taxonomic examples such as the above. Such networks are collections of nodes and directed ("is a") links representing taxonomic information. When exceptions are allowed, the network is interpreted *defeasibly*. Figure 20.2 gives a network representing this state of affairs. In such a network, if there is a link of the form $A \to B$, then information about A's is more specific than information about B's,

and hence should override it. Research on nonmonotonic inheritance focuses on the different ways in which one can make this idea precise.

The main issue in defeasible inheritance is to characterize the set of assertions that are supported by a given network. It is of course not enough to devise a representational formalism, one also needs to specify how the formalism is to be interpreted, and this is precisely the focus of much work in nonmonotonic inheritance. Such a characterization is accomplished through the notion of *extension* of a given network. There are two competing characterizations of extension for this kind of networks, one that follows the credulous strategy and one that follows the skeptical one. Both proceed by first defining the *degree* of a path through the network as the length of the longest sequence of links connecting its endpoints, and then building extensions by considering paths in ascending order of their degrees. We are not going to review the details, since many of the same issues arise in connection with default logic (which is treated to greater length below), but Horty 1994 provides an extensive survey. It is worth mentioning that since the notion of degree makes sense only in the case of acyclic networks, special issues arise when networks contain cycles (see Antonelli 1997 for a treatment of inheritance on cyclic networks).

Although the language of nonmonotonic networks is expressively limited by design (in that only links of the form "is a" can be represented in a natural fashion), such networks represent an extremely useful setting in which to test and hone one's intuitions and methods for handling defeasible information, which are then extended to more expressive formalisms. Among the latter is Reiter's "Default Logic," which is perhaps the most flexible among nonmonotonic frameworks. In Default Logic, the main representational tool is that of a *default rule*, or simply a *default*. A default is a *defeasible inference rule* of the form

$$\frac{\eta : \theta}{\xi},$$

(where η, θ, ξ are sentences in a given language, respectively called the prerequisite, the justification, and the conclusion of the default). The interpretation of the default is that if η is known, and there is no evidence that θ might be false, then the rule allows the inference of ξ. As is clear, application of the rule requires that a consistency condition be satisfied, and rules can interact in complex ways. In particular it is possible that application of a rule might cause the consistency condition to fail (as when θ is $\neg\xi$). Reiter's default logic uses the notion of an extension to make precise the idea that the consistency condition has to be met both before and after the rule is applied. Given a set Γ of defaults, an extension for Γ is, roughly, a set of defaults whose consistency condition is met both before and after their being triggered; an extension therefore represents a set of inferences that can be reasonably and consistently drawn using defaults from Γ. More in particular (and in typical circular fashion), an extension for Γ is a maximal subset Δ of Γ the conclusions of whose defaults both imply all the prerequisites of defaults in Δ and are consistent with all the justifications of defaults in Δ.

This definition can be made precise as follows. By a *default theory* we mean a pair (W, Δ), where Δ is a (finite) set of defaults, and W is a set of sentences (a world description). The idea is that W represents the strict or background information, whereas Δ specifies the defeasible information. Given a pair (T_1, T_2) of sets of sentences, a default such as the equation above is *triggered* by (T_1, T_2) if and only if $T_1 \models \eta$ and $T_2 \not\models \neg\theta$ (i.e., θ is consistent with T_2). Notice how this definition is built "on top" of $\models$: we could, conceivably, employ a different relation here. Finally we say that a set of sentences E is an *extension* for a default theory $(W; \Delta)$ if and only if

$$E = E_0 \cup E_1 \cup \ldots \cup E_n \cup \ldots,$$

where: $E_0 = W$, and

$$E_{n+1} = E_n \cup \{\xi : \frac{\eta : \theta}{\xi} \in \Delta \text{ is triggered by } (E_n, E)\}$$

(notice the occurrence of the limit E in the definition of E_{n+1}). There is an alternative characterization of extensions: given a default theory, let $\Im$ be an operator defined on sets of sentences such that for any set S of sentences, $\Im(S)$ is the

smallest set containing W, deductively closed (i.e., such that if $\mathfrak{S}(S) \vdash \phi$ then $\phi \in \mathfrak{S}(S)$), and such that if a default with consequent ξ is triggered by (S, S) then $\xi \in \mathfrak{S}(S)$. Then one can show that E is an extension for (W, Δ) if and only if E is a fixed point of $\mathfrak{S}$, i.e., if $\mathfrak{S}(E) = E$.

For any given default theory, extensions need not exist, and even when they exist, they need not be unique. Let us consider a couple of examples of these phenomena. Our first example is a default theory that has no extension: let W contain the sentence η, and let Δ comprise the single default

$$\frac{\eta : \theta}{\neg\theta}.$$

If E were an extension, then the default above would have to be either triggered or not triggered by it, and either case is impossible.

Let us now consider an example of a default theory with multiple extensions. Like before, let W contain the sentence η, and suppose Δ comprises the two defaults

$$\frac{\eta : \theta}{\neg\xi}, \quad \text{and} \quad \frac{\eta : \xi}{\neg\theta}.$$

This theory has exactly two extensions, one in which the first default is triggered and one in which the second one is. It is easy to see that at least a default has to be triggered in any extension, and that both defaults cannot be triggered by the same extension.

These examples are enough to bring out a number of features. First, it should be noted that neither one of the two characterizations of default logic given above gives us a way to "construct" extension by means of anything resembling an iterative process. Essentially, one has to "guess" a set of sentences E, and then verify that it satisfies the definition of an extension.

Further, the fact that default theories can have zero, one, or more extensions raises the issue of what inferences one is warranted in drawing from a given default theory. The problem can be presented as follows: given a default theory $(W; \Delta)$, what sentences ϕ can be regarded as *defeasible consequences* of the theory? On the face of it, there are several options available.

One option is to take the union of the extensions of the theory, and consider ϕ a consequence of a default theory (W, Δ) if and only if $\phi \in E$, for some extension E. But this option is immediately ruled out, in that it leads to endorsing contradictory conclusion, as in the second example above. It is widely believed that any viable notion of defeasible consequence for default logic must have the property that the set $\{\phi : (W, \Delta) \vdash \phi\}$ must be consistent whenever W is. Once this option is ruled out, only two alternatives are left.

The first alternative, known as the "credulous" or "bold" strategy, is to pick an extension E for the theory, and say that ϕ is a defeasible consequence if and only if $\phi \in E$. The second alternative, known as the "skeptical" or "cautious" strategy, is to endorse a conclusion ϕ if and only if ϕ is contained in *every* extension of the theory.

Both the credulous and the skeptical strategy have problems. The problem with the credulous strategy is that the choice of E is arbitrary: with the notion of extension introduced by Reiter, extensions are *orthogonal*: of any two distinct extensions, neither one contains the other. Hence, there seems to be no principled way to pick an extension over any other one. This has led a number of researcher to endorse the skeptical strategy as a viable approach to the problem of defeasible consequence. But as shown by Makinson, skeptical consequence, as based on Reiter's notion of extension, fails to be cautiously monotonic. To see this, consider the default theory (W, Δ), where W is empty, and Δ comprises the two defaults:

$$\frac{: \theta}{\theta}, \quad \text{and} \quad \frac{\theta \vee \eta : \neg\theta}{\neg\theta}.$$

This theory has only one extension, coinciding with the deductive closure of $\{\theta\}$. hence, if we put $(W, \Delta) \vdash \phi$ if and only if ϕ belongs to every extension of (W, Δ), we have $(W, \Delta) \vdash \theta$, as well as $(W, \Delta) \vdash \theta \vee \eta$ (by the deductive closure of extensions). Now consider the theory with Δ as before, but with W containing the sentence $\theta \vee \eta$. This theory has two extensions: one the same as before, but also another one coinciding with the deductive closure of $\{\neg\theta\}$, and hence not containing θ. It follows that the intersection of

the extensions no longer contains θ, so that $(\{\theta \vee \eta\}, \Delta) \not\vdash \theta$, against cautious monotony. (Notice that the same example establishes a counterexample for Cut for the credulous strategy, when we pick the extension of $(\{\theta \vee \eta\}, \Delta)$ that contains $\neg\theta$.)

It is clear that the issue of how to define a nonmonotonic consequence relation for default logic is intertwined with the way that *conflicts* are handled. The problem of course is that in this case neither the skeptical nor the credulous strategy yields an adequate relation of defeasible consequence. In Antonelli 1999 a notion of *general extension* for default logic is introduced, showing that this notion yields a well-behaved relation of defeasible consequence that satisfies all four requirements of Supraclassicality, Reflexivity, Cut, and Cautious Monotony.

A different set of issues arises in connection with the behavior of default logic from the point of view of computation. As we have seen for a given semidecidable set Γ of sentences, the set of all Γ that are a consequence of Γ in FOL is itself semidecidable. In the case of default logic, to formulate the corresponding problem one extends (in the obvious way) the notion of (semi)decidability given above to sets of defaults. The problem, then, is to decide, given a default theory (W, Δ) and a sentence ϕ whether $(W, \Delta) \vdash \phi$, where $\vdash$ is defined, say, skeptically (it doesn't really make a difference computationally whether $\vdash$ is defined skeptically or credulously). Such a problem is not even semidecidable, the essential reason being that in general, in order to determine whether a default is triggered by a pair of sets of sentences, one has to perform a consistency check. But the consistency checks are not the only source of complexity in default logic. For instance, we could restrict our language to conjunctions of atomic sentences and their negations (making consistency checks feasible). Even so, the problem of determining whether a given default theory has an extension would still be highly intractable (NP-complete, to be precise, as shown by Kautz & Selman 1991), seemingly because the problem requires checking all possible sequences of firings of defaults (see Chapter 2, Complexity, for these and related notions).

Default logic is intimately connected with certain *modal* approaches to nonmonotonic reasoning, which belong to the family of *autoepistemic logics*. Modal logics in general have proved to be one of the most flexible tools for modeling all sorts of dynamic processes and their complex interactions. Beside the applications in knowledge representation, which we are going to treat below, there are modal frameworks, known as *dynamic logics*, that play a crucial role, for instance, in the modeling of serial or parallel computation. The basic idea of modal logic is that the language is interpreted with respect to a give set of *states*, and that sentences are evaluated relative to one of these states. What these states are taken to represent depends on the particular application under consideration (they could be epistemic states, or states in the evolution of a dynamical system, etc.), but the important thing is that there are *transitions* (of one or more different kinds) between states. In the case of one transition that is both *transitive* (i.e., such that if $a \to b$ and $b \to c$ then $a \to c$) and *euclidean* (if $a \to b$ and $a \to c$ then $b \to c$), the resulting modal system is referred to as K45. Associated with each kind of state transition there is a corresponding modality in the language, usually represented as a box $\Box$. A sentence of the form $\Box A$ is true at a state s if and only if A is true at every state s' reachable from s by the kind of transition associated with $\Box$ (see Chellas 1980 for a comprehensive introduction to modal logic).

In autoepistemic logic, the states involved are epistemic states of the agent (or agents). The intuition underlying autoepistemic logic is that we can sometimes draw inferences concerning the state of the world using information concerning our own knowledge or ignorance. For instance, I can conclude that I do not have a sister given that if I did I would probably know about it, and nothing to that effect is present in my "knowledge base." But such a conclusion is defeasible, since there is always the possibility of learning new facts.

In order to make these intuitions precise, consider a modal language in which the necessity operator $\Box$ is interpreted as "it is known that." As in default logic or defeasible inheritance, the central notion in autoepistemic logic is that of an *extension* of a theory S, i.e., a consistent and self-supporting sets of beliefs that can reasonably be entertained on the basis of S. Given a set S of

sentences, let S_0 be the subset of S composed of those sentences containing no occurrences of $\Box$; further, let the *introspective closure* S_0^i of S_0 be the set

$$\{\Box\phi : \phi \in S_0\},$$

and the *negative introspective closure* S_0^n of S_0 the set

$$\{\neg\Box\phi : \phi \notin S_0\}.$$

The set S_0^i is called the introspective closure because it explicitly contains positive information about the agent's epistemic status: S_0^i expresses what is known (similarly, S_0^n contains negative information about the agent's epistemic status, stating explicitly what is not known). With these notions in place, we define an extension for S to be a set T of sentences such that:

$$T = \{\phi : \phi \text{ follows from } S \cup T_0^i \cup T_0^n \text{ in K45}\}.$$

Autoepistemic logic provides a rich language, with interesting mathematical properties and connections to other nonmonotonic formalisms. It is faithfully intertranslatable with Reiter's version of default logic, and provides a defeasible framework with well-understood modal properties.

Conclusion

There are three major issues connected with the development of logical frameworks that can adequately represent defeasible reasoning: (i) material adequacy; (ii) formal properties; and (iii) complexity. Material adequacy concerns the question of how broad a range of examples is captured by the framework, and the extent to which the framework can do justice to our intuitions on the subject (at least the most entrenched ones). The question of formal properties has to do with the degree to which the framework allows for a relation of logical consequence that satisfies the above-mentioned conditions of Supraclassicality, Reflexivity, Cut, and Cautious Monotony. The third set of issues has to do with computational complexity of the most basic questions concerning the framework.

There is a potential tension between (i) and (ii): the desire to capture a broad range of intuitions can lead to *ad hoc* solutions that can sometimes undermine the desirable formal properties of the framework. In general, the development of nonmonotonic logics and related formalisms has been driven, since its inception, by consideration (i) and has relied on a rich and well-chosen array of examples. Of course, there is some question as to whether any single framework can aspire to be universal in this respect.

More recently, researchers have started paying attention to consideration (ii), looking at the extent to which nonmonotonic logics have generated well-behaved relations of logical consequence. As Makinson (1994) points out, practitioners of the field have encountered mixed success. In particular, one abstract property, Cautious Monotony, appears at the same time to be crucial and elusive for many of the frameworks to be found in the literature. This is a fact that is perhaps to be traced back, at least in part, to the above-mentioned tension between the requirement of material adequacy and the need to generate a well-behaved consequence relation.

The complexity issue appears to be the most difficult among the ones that have been singled out. Nonmonotonic logics appear to be stubbornly intractable with respect to the corresponding problem for classical logic. This is clear in the case of default logic, given the ubiquitous consistency checks. But beside consistency checks, there are other, often overlooked, sources of complexity that are purely combinatorial. Other forms of nonmonotonic reasoning, beside default logic, are far from immune from these combinatorial roots of intractability. Although some important work has been done trying to make various nonmonotonic formalisms more tractable, this is perhaps the problem on which progress has been slowest in coming.

References

Antonelli, G. Aldo. 1997. "Defeasible inheritance over cyclic networks." *Artificial Intelligence* 92(1): 1–23. [A treatment of nonmonotonic inheritance

networks that include cycles, inspired by Kripke's 3-valued approach to truth theories.]

——. 1999. "A directly cautious theory of defeasible consequence for default logic via the notion of general extension." *Artificial Intelligence* 109 (1–2): 71–109. [Introduces a well-behaved consequence relation for default logic, based on a generalization of Reiter's original notion of extension.]

Börger, Egon, Grädel, Erich, and Gurevich, Yuri. 1997. *The Classical Decision Problem*. Berlin and New York: Springer-Verlag. [A standard reference on the complexity of the decision problem for FOL and many of its fragments.]

Chellas, Brian F. 1980. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press. [The standard introduction to modal logic. Cover standard possible-world semantics as well as other variants.]

Ebbinghaus, H.-D., Flum, J., and Thomas, W. 1994. *Mathematical Logic*, 2nd ed. New York and Berlin: Springer-Verlag. [A graduate-level introduction to symbolic logic. Notable for its treatment of Lindström's results.]

Enderton, Herbert. 1972. *A Mathematical Introduction to Logic*, 2nd ed. New York: Academic Press. New York: Harcourt/Academic Press. [An introduction to logic aimed at advanced undergraduates. Enderton's axiomatization of FOL is widely used.]

Fitting, Melvin. 1990. *First-order Logic and Automated Theorem Proving*. New York and Berlin: Springer-Verlag. [An introduction to formal logic emphasizing the mechanization of theorem-proving. Notable for its treatment of both resolution and tableaux methods.]

Gabbay, D. M., Hogger, C. J., and Robinson, J. A., eds. 1994. *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 3. Oxford: Oxford University Press. [Contains excellent surveys of all the major nonmonotonic formalisms as well as articles addressing the foundations of nonmonotonic logic.]

Gallier, Jean H. 1986. *Logic for Computer Science: Foundations of Automated Theorem Proving*. New York: Harper & Row. [A technically sophisticated treatment of first-order logic using mainly proof-theoretic methods. Covers Cut-elimination for Gentzen systems, resolution, and many-sorted first-order logic.]

Galton, Antony. 1990. *Logic for Information Technology*. Chichester and New York: John Wiley & Sons. [A thorough introduction to proof systems for first-order logic and their metatheory.]

Ginsberg, M. L., ed. 1987. *Readings in Nonmonotonic Reasoning*. Los Altos, CA: Morgan Kauffman. [A collection of early papers in nonmonotonic logic. Somewhat hard to find, but invaluable.]

Gödel, Kurt. 1930. "Die Vollständigkeit der Axiome des logischen Funktionenkalküls." *Monatshefte für Mathematik und Physik* 37. [A classic.]

Horty, John F. 1994. "Some direct theories of nonmonotonic inheritance." In Gabbay et al. 1994: 111–87. [A very clear presentation of all the major issues in defeasible inheritance.]

——. 2002. "Skepticism and floating conclusions." *Artificial Intelligence Journal* 135(1–2): 55–72. [This article addresses some issues in the foundations of defeasible reasoning. Makes the case that, contrary to what many have argued, floating conclusions are not always warranted.]

Kautz, H. and Selman, B. 1991. "Hard problems for simple default logic." *Artificial Intelligence Journal* 49: 243–79. [A seminal work addressing complexity issues in nonmonotonic inference.]

Lindström, Per. 1969. "On extensions of elementary logic." *Theoria* 35. [The original reference for Lindström's theorems.]

Makinson, David. 1994. "General patterns in nonmonotonic reasoning." In Gabbay et al. 1994: 35–110. [Best general treatment of the issues concerning abstract consequence relations for various defeasible formalisms.]

Stalnaker, Robert. 1994. "Nonmonotonic consequence relations." *Fundamenta Informatica* 21: 7–21. [A deep paper assessing several abstract properties of nonmonotonic consequence relations.]

Tarski, Alfred. 1935. "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Logica*, pp. 261–405: English tr. in Tarski 1956: 152–278. [The classic, original treatment of semantics for FOL.]

——. 1956. *Logic, Semantics, and Metamathematics*. Oxford: Oxford University Press, ed. and tr. J. H. Woodger. [A collection of Tarski's mostly technical papers.]

Chapter 21

Probability in Artificial Intelligence

Donald Gillies

Introduction

The first two sections of this chapter provide an account of how the development of artificial intelligence (AI) led to some remarkable innovations in probability theory – most notably to the new theory of Bayesian networks. The final section presents some discussion of the philosophical problems posed by these new ideas. As we shall see, they have a considerable impact on quite a number of traditional questions in the philosophy of science.

The Breakthrough with Expert Systems in the 1970s

Research in AI began in the 1950s and many important ideas were developed by the pioneers (see Chapter 9, The Philosophy of AI and its Critique). Then in the 1970s a breakthrough was produced by the creation of expert systems. The lead here was taken by the Stanford heuristic programming group, particularly Buchanan, Feigenbaum, and Shortliffe. What they discovered was that the key to success was to extract from an expert the knowledge he or she used to carry out a specialized task, and then code this knowledge into the computer. In this way they were able to produce "expert systems" which performed specific tasks at the level of human experts. One of the most important of these early expert systems (MYCIN) was concerned with the diagnosis of blood infections. This system will now be briefly described, and it will then be shown that its implementation led to the problem of how to introduce probability into AI.

MYCIN was developed in the 1970s by Edward Shortliffe and his colleagues in collaboration with the infectious diseases group at the Stanford medical school. The medical knowledge in the area was codified into rules of the form: IF such and such symptoms are observed, THEN likely conclusion is such and such. MYCIN's knowledge base comprised over 400 such rules which were obtained from medical experts. An example of such a rule will be given in a moment, but first it would be as well to present some evidence of MYCIN's success.

To test MYCIN's effectiveness a comparison was made in 1979 of its performance with that of 9 human doctors. The program's final conclusions on 10 real cases were compared with those of the human doctors, including the actual therapy administered. Eight other experts were then asked to rate the 10 therapy recom-

mendations and award a mark, without knowing which, if any, came from a computer. They were requested to give 1 for a therapy which they regarded as acceptable and 0 for an unacceptable therapy. Since there were 8 experts and 10 cases, the maximum possible mark was 80. The results were as follows (cf. Jackson 1986: 106):

MYCIN	52	Actual therapy	46
Faculty-1	50	Faculty-4	44
Faculty-2	48	Resident	36
Inf dis fellow	48	Faculty-5	34
Faculty-3	46	Student	24

So MYCIN came first in the exam, though the difference between it and the top human experts was not significant.

Let us now examine one of MYCIN's rules. The following rule is given by Shortliffe and Buchanan in their 1975, p. 357.

> *If*: (1) the stain of the organism is gram positive (S_1), and
> (2) the morphology of the organism is coccus (S_2), and
> (3) the growth conformation of the organism is chains (S_3)
>
> *Then*: there is suggestive evidence (0.7) that the identity of the organism is streptococcus (H_1).

In symbols this could be written: If S_1 & S_2 & S_3, then there is suggestive evidence p that H_1, where $p = 0.7$. Here S_1, S_2, S_3 are the observations/symptoms, which support hypothesis H_1 to a particular degree. These rules were obtained from the medical experts. The numbers they contain such as 0.7 look like probabilities, and they too were obtained from the experts. The expert was in fact asked: "On a scale of 1 to 10, how much certainty do you affix to this conclusion?" The answer was then divided by 10.

It looks as if Shortliffe and Buchanan are using probability in the subjective sense to measure the degree of personal belief held by an expert. This at once raises the question of why subjective probabilities (see the appendix to this chapter) obtained from experts are preferred to objective probabilities (see appendix) obtained from data. Shortliffe and Buchanan do consider this question, and they answer (1975: 352–5) that in typical medical applications there is not enough data to obtain the requisite objective probabilities. This in turn is because of the inadequacy of hospital records, and the changes which are continually occurring in disease categories. It is interesting to note that only three years previously, another group working on computer diagnosis had reached exactly the opposite conclusion. This research group, working in Leeds, was headed by de Dombal. Their results are contained in de Dombal et al. 1972, and Leaper et al. 1972. They will now be briefly summarised.

De Dombal's group created a computer diagnosis system for acute abdominal pain based on Bayesian reasoning (see appendix), but using objective probabilities obtained from a sample of 600 patients. In a test involving 472 patients, this system was correct in 91.1 percent of cases, while the clinical team was correct in 79.7 percent of cases. The system was then changed by using instead of these objective probabilities, subjective probabilities obtained from the clinicians. Its performance for the same 472 patients, but using these subjective probabilities, dropped from 91.1 to 82.2 percent. Moreover the performance of the computer system with subjective probabilities was actually worse than that of the clinicians in the case of diseases which occurred relatively rarely. The conclusion seemed inescapable that human doctors are bad at estimating probabilities, especially in the case of diseases which occur infrequently, and that therefore the use of objective probabilities obtained from data is preferable.

Despite the strong evidence for this conclusion, it was ignored for the next 20 years, and nearly every researcher in the field made use of subjective probabilities. There are two possible reasons for this. First of all it may, in many cases, have been difficult to obtain objective probabilities from data. Secondly the general methodology of expert systems research, since it involved obtaining knowledge from the experts, may have encouraged the idea of obtaining probabilities as the degrees of belief of these experts. I will discuss further the question of objective versus subjective probabilities in the final section of this chapter, which deals with the related philosophical problems. Let us now return to a

consideration of MYCIN, and the sample rule given earlier.

So far we have rather assumed that the number 0.7 in the rule from MYCIN is an ordinary probability, but this is not the case, as Shortliffe and Buchanan make clear in the following passage (1975: 358):

> this rule at first seems to say $P(H_1 \mid S_1 \& S_2 \& S_3) = 0.7, \ldots$. Questioning of the expert gradually reveals, however, that despite the apparent similarity to a statement regarding a conditional probability, the number 0.7 differs significantly from a probability. The expert may well agree that $P(H_1 \mid S_1 \& S_2 \& S_3) = 0.7$, but he becomes uneasy when he attempts to follow the logical conclusion that therefore $P(\text{not}.H_1 \mid S_1 \& S_2 \& S_3) = 0.3$. The three observations are evidence (to degree 0.7) *in favor* of the conclusion that the organism is a streptococcus and should not be construed as evidence (to degree 0.3) *against* streptococcus.

Shortliffe and Buchanan used this observation to motivate the introduction of a *nonprobabilistic* model of evidential strength. Their measure of evidential strength was called a *certainty factor*, and certainty factors neither obeyed the standard axioms of probability theory, the Kolmogorov axioms (see appendix), nor combined like probabilities.

Certainty factors were criticized by those who favored a probabilistic approach – cf. Adams 1976 and Heckerman 1986 – and in fact the next expert system we will consider (PROSPECTOR) did move more in the direction of standard probability.

PROSPECTOR, an expert system for mineral exploration, was developed in the second half of the 1970s at the Stanford Research Institute. A good general account of the system is given by Gaschnig in his 1982. PROSPECTOR's most important innovation was to represent knowledge by an *inference network* (or *net*). This is motivated by Duda et al in their 1976 as follows (p. 1076):

> A collection of rules about some specific subject area invariably uses the same pieces of evidence to imply several different hypotheses. It also frequently happens that several alternative pieces of evidence imply the same hypothesis. Furthermore, there are often chains of evidences and hypotheses. For these reasons it is natural to represent a collection of rules as a graph structure or *inference net*.

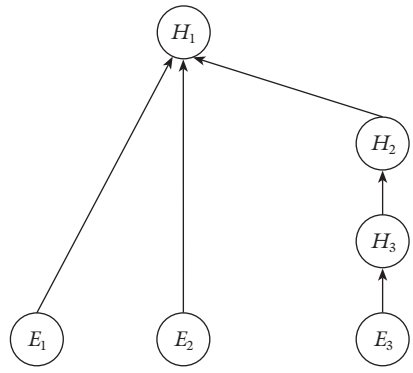

Figure 21.1: Part of PROSPECTOR's inference network. H_1 = There are massive sulfide deposits. H_2 = There are clay minerals. H_3 = There is a reduction process. E_1 = Barite is overlying sulfide. E_2 = Galena, sphalerite, or chalcopyrite fill cracks in rhyolite or dacite. E_3 = There are bleached rocks.

A part of PROSPECTOR's inference network is shown in figure 21.1. Evidence E_1 is taken as supporting hypothesis H_1, and this is indicated by the arrow joining them in the inference network. Similarly E_2 supports hypothesis H_1, while E_3 supports H_3 which supports H_2 which supports H_1. Note how these rather complicated relations are simply and elegantly represented by the arrows of the network. Each inference arrow has a strength associated with it, and this is obtained from the expert as in the case of MYCIN.

PROSPECTOR, however, differs from MYCIN in using subjective Bayesianism (see appendix) rather than certainty factors. This subjective Bayesianism is not entirely pure, since, as in the case of MYCIN, it is combined with the use of fuzzy logic formulae (see appendix). This use of fuzzy logic tended to disappear in further developments.

In PROSPECTOR, Bayesianism is formulated using odds rather than probabilities. The odds on a hypothesis $H[O(H)]$ are defined as follows:

$$O(H) = P(H)/P(\neg H)$$

Writing down Bayes theorem (see appendix) first for H and then for $\neg H$, we get

$$P(H \mid E) = P(E \mid H)P(H)/P(E)$$

$$P(\neg H \mid E) = P(E \mid \neg H)P(\neg H)/P(E)$$

So dividing gives

$$O(H \mid E) = \lambda(E)O(H) \qquad (21.1)$$

where λ(E) is the likelihood ratio $P(E \mid H)/P(E \mid \neg H)$. (21.1) is the odds and likelihood form of Bayes theorem, and it is used in PROSPECTOR to change the prior odds on H to the posterior odds given evidence E.

Let us now consider the problems which arise if we have several different pieces of evidence $E_1, E_2, \ldots, E_n$ say. We might in practice have to update using any subset of these pieces of evidence $E_i, E_j, \ldots, E_k$ say, where $(i, j, \ldots k)$ is any subset of $(1, 2, \ldots, n)$. If we use (21.1), this would involve having values of $\lambda(E_i \,\&\, E_j \,\&\, \ldots \,\&\, E_k)$ for all subsets of $(1, 2, \ldots n)$. When we remember that, on this approach the values of λ are obtained from the domain experts, we can see that obtaining the requisite values of λ is scarcely possible. Clearly some simplifying assumptions are necessary to produce a workable system, and the designers of PROSPECTOR therefore made the following two conditional independence (see appendix) assumptions:

$$P(E_1, \ldots, E_n \mid H) = P(E_1 \mid H) \ldots P(E_n \mid H) \qquad (21.2)$$

$$P(E_1, \ldots, E_n \mid \neg H) = P(E_1 \mid \neg H) \ldots P(E_n \mid \neg H) \qquad (21.3)$$

Given these assumptions, the whole problem of updating with many pieces of evidence becomes simple, and, in fact,

$$O(H \mid E_1 \,\&\, \ldots \,\&\, E_n) = \lambda_1, \lambda_2 \ldots \lambda_n \; O(H)$$
where $\lambda_i = \lambda(E_i)$

The only remaining problem was whether the conditional independence assumptions (21.2) and (21.3) are plausible. The search for a justification of these assumptions led, as we shall see in the next section, to the modification of the concept of inference network, and the emergence of the concept of *Bayesian network*.

The Emergence of Bayesian Networks in the 1980s

The concept of Bayesian networks was introduced and developed by Pearl in a series of papers: Pearl 1982, 1985a, 1985b, 1986, Kim & Pearl 1983, and a book: Pearl 1988. An important extension of the theory was carried out by Lauritzen and Spiegelhalter (1988), while Neapolitan's 1990 book gave a clear account of these new ideas and helped to promote the use of Bayesian networks in the AI community. In what follows, I will comment on a few salient features of Bayesian networks which will be important when we consider their philosophical implications in the next section.

The actual term *Bayesian* (or *Bayes*) *network* was introduced in Pearl's 1985b, where it is defined as follows (1985b: 330):

> Bayes Networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by conditional probabilities.

This verbal account is illustrated by a diagram which is reproduced in figure 21.2.

If we compare the network of figure 21.2 with that of figure 21.1, two differences should be noted immediately. First of all, the arrows in the inference network of figure 21.1 represent a relation of support holding between e.g. E_3 and H_3, while the arrows in the Bayesian network of figure 21.2 represent causal influences, so that, e.g. the arrow joining X_1 to X_2 means that X_1 causes X_2. Secondly, corresponding to the first difference, we can say that, in a certain sense, the arrows of a Bayesian network run in the opposite direction to those of an inference network. Pearl puts this point as follows (1986: 253–4):

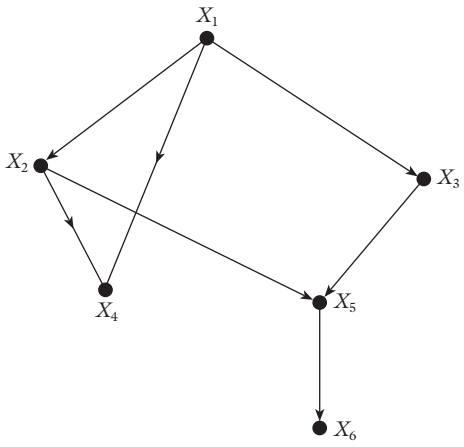

Figure 21.2: A Bayes network

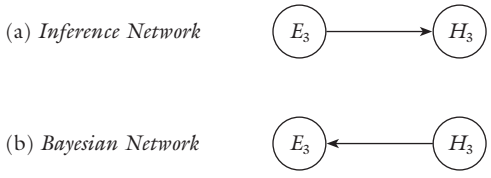

Figure 21.3: A set of nodes from figure 21.1

> in many expert systems (e.g. MYCIN), . . . rules point from evidence to hypothesis (e.g. if symptom, the disease), thus denoting a flow of mental inference. By contrast, the arrows in *Bayes'* networks point from causes to effects or from conditions to consequence, thus denoting a flow of constraints in the physical world.

This reversal of arrows from inference networks to Bayesian networks is illustrated in figure 21.3, which shows one pair of nodes taken from the portion of PROSPECTOR's inference network shown in figure 21.1.

In figure 21.3, E_3 = There are bleached rocks, while H_3 = There is a reduction process. From the point of view of an inference network (a), we regard the evidence of bleached rocks as supporting the hypothesis that there is a reduction process, while, from the point of view of a Bayesian network (b), we regard there being a reduction process as the cause of there being bleached rocks. In his 1993, Pearl gives an account of his discovery of Bayesian networks,

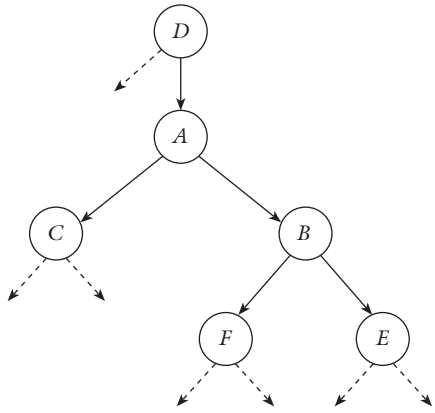

Figure 21.4: Pearl's (1982) tree example

and says that one factor that led him to the idea was his consideration of the concept of influence diagrams introduced by Howard and Matheson (1984). Pearl decided to limit the influences specifically to causal influences.

But what is the advantage of this reversal of arrows and introduction of causal links? To answer this question, we must return to the question of the conditional independence assumptions which are needed in order to make Bayesian updating (see appendix) feasible. Before doing do, however, I would like to make one further point about Bayesian networks. If, in such a network, an arrow runs from node A to node B, then A is said to be a parent of B, and B a child of A. If a node has no parents, it is called a root, so that in figure 21.2, X_1 is a root. In a Bayesian network, it is possible for a child to have several parents. Thus in figure 21.2, X_5 has parents X_2 and X_3. If, however, every child has at most one parent, the network is called a tree. Pearl started his investigation of networks with trees since they are mathematically simpler. Let us similarly begin our account of the conditional independence assumptions in the case of trees by considering Pearl's first paper on the subject, his 1982. This paper is particularly helpful in showing how the method of Bayesian networks developed from PROSPECTOR's method of inference networks.

In his 1982, Pearl considers the example of a tree illustrated in figure 21.4. Here A, B, C, . . . are variables standing for hypotheses or observations. A takes the values A_1, A_2, . . . , and

similarly for the other variables. Pearl now introduces the conditional independence assumptions which he is going to make as follows (1982: 134):

> let $D_d(B)$ stand for the data obtained from the tree rooted at B, and $D^u(B)$ for the data obtained from the network above B. The presence of only one link connecting $D^u(B)$ and (B) implies:
>
> $$P(B_j \mid A_i, D^u(B)) = P(B_j \mid A_i) \ldots$$

In other words, a node is conditionally independent given its parent of the rest of the network except its descendents. If we here substitute "parents" for "parent," we get a statement of the conditional independence assumptions which define a Bayesian network.

In fact Bayes's theorem, together with the above conditional independence assumptions, yields the product rule:

$$P(B_i \mid D^u(B), D_d(B)) = \alpha P(D_d(B) \mid B_i) \cdot P(B_i \mid D^u(B)) \qquad (21.4)$$

where α is a normalization constant. Pearl now makes a most interesting comparison between (21.4) and the updating formula of PROSPECTOR given as (21.1) above.

$$O(H \mid E) = \lambda(E)O(H) \qquad (21.1)$$

Since $D^u(B)$, $D_d(B)$ represents the total data, or evidence E bearing on our hypothesis B_i, $P(B_i \mid D^u(B), D_d(B))$ corresponds to $O(H \mid E)$. $P(D_d(B) \mid B_i)$ has the form of a likelihood (see appendix) as does $\lambda(E)$; while $P(B_i \mid D^u(B))$ corresponds to the prior probability of traditional Bayesianism in the following sense (Pearl 1982: 134):

> the multiplicative role of the prior probability . . . is taken over by the conditional probability of a variable based *only* on the evidence gathered by the network *above* it, excluding the data collected from below. . . . The root is the only node which requires a prior probability estimation, since it has no network above. $D^u(B)$ should be interpreted as the available background knowledge which remains unexplicated by the network below. This interpretation renders $P(B_i \mid D^u(B))$ identical to the classical notion of subjective prior probability.

These analogies can be further emphasized by the notation

$$\lambda(B_i) =_{\text{def}} P(D_d(B) \mid B_i)$$

$$\pi(B_i) =_{\text{def}} P(B_i \mid D^u(B))$$

which allows one to write (21.4) in the form

$$P(B_i \mid E) = \lambda\alpha(B_i)\pi(B_i). \qquad (21.5)$$

The analogy between (21.1) and (21.5) is clear.

On the basis of (21.5), Pearl develops an algorithm which allows Bayesian updating to take place. If one of the variables which represents an observation is set to a particular value, the changes brought about by this new information in all the probabilities throughout the tree can be computed in an efficient manner. In subsequent work he extends this updating algorithm to more complicated networks. Kim & Pearl 1983 generalizes from trees to Bayesian networks which are singly connected, i.e. there exists only one (undirected) path between any pair of nodes. Pearl in his 1986 tackled the further extension to Bayesian networks which are multiply connected. This problem was also investigated by Lauritzen & Spiegelhalter who in their 1988 solved it using the idea of reducing a multiply connected network to a tree of cliques. Their algorithm has been generally adopted by the AI community.

Let us now turn from these powerful mathematical developments to the consideration of a conceptual point. It will have been noted that two rather different definitions of Bayesian network have been given. The first definition is in terms of causes. Thus in figure 21.2 the arrows are taken as denoting a causal link between the two nodes which they join. The second definition is by contrast purely probabilistic. In figure 21.2 the variables $X_1, X_2, \ldots X_6$ are taken to be random variables with a joint probability distribution (see appendix), and the network becomes a Bayesian network if the relevant conditional

independence assumptions are satisfied. I will henceforth use the term "Bayesian network" for networks defined purely probabilistically in the manner just explained, and call the networks defined in terms of causes: "causal networks." Pearl tends, however, to use the terms "Bayesian network" and "causal network" interchangeably, because he believes the two notions to be closely connected. More specifically, his idea is that if in a network the parents of every node represent the direct causes of that node, then the relevant conditional independence assumptions will automatically be satisfied. As he says (1993: 52):

> Causal utterances such as "*X* is a direct cause of *Y*" were given a probabilistic interpretation as distinctive patterns of conditional independence relationships that can be verified empirically.

A suggested link between causality and conditional independence in fact goes back to Reichenbach 1956. Reichenbach considers two events, *B* and *C* say, which are correlated. For example, in a traveling troupe of actors, *B* = the leading lady has a stomach upset, and *C* = the leading man has a stomach upset. We can explain such correlations, according to Reichenbach, by finding a common cause, namely that the leading lady and the leading man always have dinner together. The common stomach upsets occur when the food in the local restaurant has gone off. Denote "dining together" by *A*. We then have the causal graph shown in figure 21.5.

Reichenbach then claimed that, conditional on *A*, *B*, and *C* were no longer correlated but independent, i.e. $P(B\&C \mid A) = P(B \mid A)P(C \mid A)$. He also expressed this idea by saying that a common cause *A* screens one of its effects *B* off from the other *C*. Reichenbach's causal fork is just a simple case of a Bayesian network. We can indeed apply his term "screening off" to Bayesian networks by saying that in such networks, the parents of a node screen it off from all the other nodes in the network except its descendents.

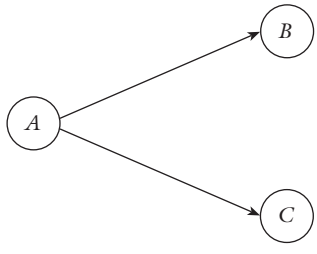

Figure 21.5: Dining together

We are now in a position to summarize the ingenious way in which Bayesian networks solved the problem of handling uncertainty in expert systems. In most of the domains considered, e.g. medical diagnosis, a domain expert is very familiar with the various causal factors operating. It should therefore be an easy matter to get him or her to provide a causal network. By the addition of probabilities this can be turned into a Bayesian network. In earlier systems such as MYCIN or PROSPECTOR, conditional independence assumptions were made for the purely *ad hoc* and pragmatic reason of allowing the updating to become possible. For Bayesian networks, however, the causal information obtained from the expert provides a justification for making a set of conditional independence assumptions in the manner first suggested by Reichenbach. Moreover as Pearl, Lauritzen, and Spiegelhalter have shown, this set of conditional independence assumptions is sufficient to allow Bayesian updating to become computationally feasible. Everything fits together in a most satisfying manner. There is only one weak link in the chain. It turns out, as we shall see in the next section, that it is possible to have a *bona fide* causal network in which the requisite conditional independence assumptions are not satisfied.

Philosophical Problems Connected with Probability in AI

The preceding sections have outlined some remarkable developments in AI. Let us now turn to a consideration of the philosophical implications of these developments. In fact they have a profound impact on at least three central problems in the philosophy of science. The first of these is the Bayesianism versus non-Bayesianism debate. This has continued among philosophers of science for the last 50 years with no signs of abating. In the 1950s the major contenders were

Carnap (in favor of Bayesianism) versus Popper (against Bayesianism). In the late 1980s and 1990s, we have had Howson & Urbach (1989) in favor of Bayesianism, and Miller (1994) against, while the most recent developments in the debate as seen by leading experts in the area are to be found in Corfield & Williamson 2001. The new results in AI are clearly relevant to this controversy, and indeed would seem to favor the Bayesian camp, though, as we shall see, this support is more qualified than it might at first appear.

Secondly, and closely connected, there is the problem of the interpretation of probability. Here the main division is between those who favor an objective interpretation, such as frequency or propensity, and those who favor a subjective, or degree of belief, interpretation. Of course pluralist positions are also possible in which different interpretations are considered appropriate in different contexts. One such pluralist position is defended in Gillies 2000, chs. 8 and 9. Once again the developments in AI just described are clearly relevant to this issue, and indeed would seem to favor the subjective interpretation of probability.

Thirdly there is the problem of the relation of causality and probability. Of course the problem of causality is one of the oldest and most central questions in Western philosophy. Aristotle, Hume, and Kant all made fundamental contributions to the analysis of causality. In the last 50 years, however, the debate has taken a new turn through the emergence of a notion of indeterminate causality (see appendix), and the corresponding investigation of the relations between causality and probability. Many leading philosophers of science, including Cartwright, Fetzer, Popper, Reichenbach, Salmon, and Suppes, have written on this issue, and a good recent survey of these philosophical developments is to be found in Salmon 1998. Now clearly the new theory of Bayesian networks, involving as it does a novel combination of causality and probability, is highly relevant to this debate. These implications of AI for the philosophy of science are so important that considerations of AI are bound to play a role of increasing importance in philosophy of science in the coming years. In this short chapter, there is room only for a few preliminary observations, and I will confine myself to discussing just the first two of the above problems.

Let us start therefore with the long-running controversy about Bayesianism. As so often happens in such controversies, it turns out that the definition of Bayesianism is not entirely clear, and I believe that the controversy involves two rather different issues. The first of these issues is the question of whether we should use the standard mathematical calculus of probability in handling uncertainty, or whether some other calculus might be appropriate. Here, of course, the Bayesians favor the use of the standard calculus. As an example of a non-Bayesian position we can take the view of Popper (see his 1934, and, for a discussion, Gillies 1998a) that the corroboration of universal laws of science [$C(H, E)$] is not a probability function, i.e. does not satisfy the standard axioms of probability. In symbols the claim is that $C(H, E) \neq P(H \mid E)$.

As we have seen, this debate occurred also in the AI context. MYCIN used a nonprobabilistic measure of evidential strength, and several other nonprobabilistic approaches were proposed and developed by AI workers. (For some details, see Ng & Abramson 1990.) However, the development of AI has given a relatively unequivocal verdict. Probabilistic measures have proved much more successful in practice than nonprobabilistic measures, and the latter have tended to disappear. AI has thus supported Bayesianism in this first sense. It should be added, however, that this does not give a decisive verdict against Popper's ideas on corroboration. Popper was considering the corroboration of hypotheses which were universal scientific laws. Most AI systems, however, have as hypotheses singular statements, such as "this patient's infection is caused by streptococci" or "that mountain range contains massive sulfide deposits." It is possible that Bayesianism is appropriate for singular statements, while a non-Bayesian approach is appropriate for universal hypotheses (see Gillies 1998a: 154–5 for arguments in favor of this position).

Let us now turn to the second and rather different issue involved in the Bayesian controversy. It can be most easily approached by considering the form that the debate has taken within statistics. Classical statisticians such as Neyman were strongly opposed to Bayesianism. Yet Neyman never used any formal system other than the standard mathematical theory of probability.

Neyman was clearly not an anti-Bayesian in the sense we have just considered. In what sense, then, was he against Bayesianism? The answer is not immediately clear, since, because he accepted standard probability theory, Neyman *a fortiori* accepted Bayes' theorem. The answer to this conundrum is that the second issue in the Bayesian controversy is really about the interpretation of probability. Neyman, following von Mises, regarded the objective interpretation of probability as the only valid one (cf. Gillies 1998b: 6–9). This meant that some applications of Bayes's theorem were illegitimate in his eyes because they necessitated giving a degree of belief interpretation to some of the probabilities used. This applied particularly to the case of giving an *a priori* distribution to a fixed, but unknown, parameter θ. Since θ is fixed and does not vary randomly, it does not make sense to assign it an objective probability distribution, but, since it makes perfect sense for someone to have different degrees of belief in different possible values of θ, θ can easily be given a subjective probability distribution. Many Bayesian analyses involve giving *a priori* distributions to parameters such as θ, and so become illegitimate to a strict objectivist such as Neyman. To sum up: the second issue involved in the Bayesian controversy is really the same as our second general philosophical question concerning the interpretation of probability.

What have the AI developments given above shown as regards this controversy? It is immediately clear that they have lent support to the subjective interpretation of probability. Pearl has always argued for a subjective degree of belief interpretation of the probabilities in Bayesian networks, and this remains true of his latest, highly interesting, paper on the foundations of the subject (Pearl 2001). In this paper he describes himself as "only a half-Bayesian." However, his departure from standard Bayesianism arises because he thinks that prior probability distributions are inadequate to express background knowledge, and that one needs also to use causal judgments which cannot be expressed in probabilistic terms. As far as the interpretation of probability is concerned he remains faithful to the subjective, degree-of-belief, view which he says he adopted in 1971 after reading Savage.

Lauritzen and Spiegelhalter were also working in the tradition of subjective Bayesianism, but they seem less definitely committed to this view than Pearl. This is what they say (1988: 159):

> Our interpretation of probabilities is that of a subjectivist Bayesian . . . This seems a convenient and appropriate view in an area concerned with the rational structuring and manipulation of opinion, and the subjectivist objectives of a coherent system of probabilities representing belief in verifiable propositions, successively updated on the basis of available evidence, appears to fit remarkably the objectives of expert systems research. However, many of the techniques presented here are appropriate in disciplines where graphical structures are used and a frequentist interpretation is more appropriate, such as complex pedigree analysis in genetics.

So Lauritzen and Spiegelhalter think that in some cases at least the probabilities in Bayesian networks might be given an objective interpretation. Neapolitan (1990) is also favorable to objective probabilities in Bayesian networks. So although Bayesian networks were created within the tradition of subjective Bayesianism, it might nonetheless be possible to interpret the probabilities they contain objectively. Arguably this is likely to be a good strategy in many cases.

A first argument in favor of an objective interpretation is an appeal to the results, given earlier, of de Dombal's group at Leeds. Their test showed that a diagnostic computer system performed far better when using objective probabilities derived from data than when it used subjective probabilities obtained from the clinicians.

These results are very striking, but the issue is not simply one of a choice between different ways of interpreting probabilities. It should now be pointed out that this choice carries with it methodological implications. If we are interpreting the probabilities as objective, then any proposed value of a probability must be seen as a conjecture which could be right or wrong, and may therefore need testing. Thus objective probabilities lead to a Popperian methodology of conjectures and refutations in which testing plays a central role. This is indeed the methodology of classical statistics.

Let us next contrast this with the use of subjective probabilities. Any such probability expresses the degree of belief of an individual at a particular moment. Further evidence does not refute the claim that that individual held that degree of belief at that time. It may however lead the individual to change his or her degree of belief in the light of the new evidence. In the Bayesian approach, the belief change takes place through Bayesian conditionalization or updating (see appendix), i.e. through the change from a prior probability $P(H)$ to a posterior probability $P(H \mid E)$. To sum up then. The use of objective probabilities goes with the Popperian methodology of statistical testing; while the use of subjective probabilities goes with the methodology of Bayesian conditionalization. There have been examples which show that the use of a testing methodology can be advantageous in the construction of Bayesian networks.

One such example (see Sucar 1991, and Sucar, Gillies, & Gillies 1993) concerned a medical instrument called an "endoscope." This allowed a doctor to put into the colon of a patient a small camera which transmitted an image of the interior of the colon to a television screen. In this image an expert could recognize various things in the interior of the colon. Let us take two such things as examples. One is called the "lumen," which is the opening of the colon. Despite its name, it generally appeared as a large dark region; but sometimes it was smaller and surrounded by concentric rings. Another is called a "diverticulum" and is a small malformation in the wall of the colon, which can cause some illnesses. A diverticulum generally appeared as a dark region, smaller than the lumen, and often circular. It was a problem then to program a computer to recognize from the image the lumen or a diverticulum. This is a typical problem of computer vision. To solve it, an attempt was made to construct a Bayesian network with the help of an expert in medical endoscopy.

Figure 21.6 shows only a small part of this network, but it is sufficient to illustrate the points which are to be made. *L* stands for the lumen which causes a large dark region (LDR) to appear on the screen. This in turn produces values for the variables *S* (size of the region), and *M* and *V* (mean and variance of the light intensity of the region).

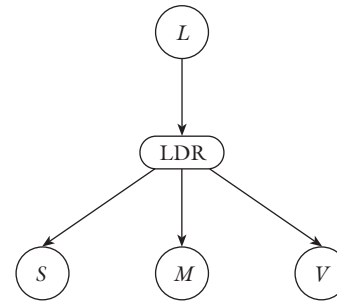

Figure 21.6: Endoscopic Bayesian network

In every Bayesian network certain assumptions of independence or conditional independence are made. In this case, *S*, *M*, *V* must be conditionally independent of *L* and mutually conditionally independent, given LDR. Using a Popperian testing methodology, these assumptions were considered as conjectures which needed to be checked by statistical tests. These tests showed, however, that the conditional independence assumptions were not satisfied. In fact it turned out that, given LDR, *M* and *V* were strongly correlated rather than independent.

The response to this situation was to eliminate one of the two parameters *M* and *V* on the grounds that, since they were correlated, only one could give almost as much information as both. The results of this elimination were tested using a random sample of more than 130 images of the colon. It turned out that the elimination of one of the parameters gave better results than those obtained using all three parameters. For example, using all three parameters (*S*, *M*, *V*), the system recognized the lumen correctly in 89 percent of cases, while, if it was modified by eliminating *M*, and using only (*S*, *V*) it recognized the lumen correctly in 97 percent of cases (for further details, see Sucar, Gillies, & Gillies 1993: 206). At first sight this seems a paradox, because these better results were obtained using less information. The explanation is simple, however. Undoubtedly there is more information in all three parameters (*S*, *M*, *V*) than in only two (*S*, *V*). But the greater amount of information in the three parameters was used with mathematical assumptions of conditional independence which were not correct. The lesser amount of

information in the two parameters *S*, *V* was, by contrast, used with true mathematical assumptions. So less information in a correct model worked better than more information in a mistaken model. Moreover, since the modified Bayesian network was simpler, the calculations using it were carried out more quickly. So, to conclude, the modified Bayesian network was more efficient, and gave better results. This shows the value of using objective probabilities, and a Popperian methodology of statistical testing.

The example also shows that it is possible to obtain a causal network from an expert for which the assumptions of conditional independence are not satisfied. Thus, although causal networks are useful heuristic guides for the construction of Bayesian networks, they are not infallible guides. This raises the third of the philosophical questions mentioned above, namely the question of how causality is related to probability. However, space unfortunately does not permit a discussion of this question here.

I conclude by briefly mentioning some further investigations into the use of probability in AI, which raise important philosophical questions. This chapter has focused on the so-called symbolic approach to AI, but there is in addition the approach using neural networks which also makes extensive use of probability. An account of some of the problems here is to be found in Williams 2001. Another area of AI involving probability is machine learning, that is to say, the attempt to program computers to induce laws from data. An account of how new results in machine learning impinge upon longstanding philosophical discussions of induction is to be found in Gillies 1996.

Appendix

In the chapter some technical terms from logic, probability, and causality are used. The meaning of these terms is explained in what follows. The terms defined are underlined, and the text contains a reference to this appendix when the term is first introduced.

Logic is concerned with propositions such as *A* = Jones is bald. The negation of a proposition, e.g. Jones is not bald, is written as not.*A*, or $\neg A$. In standard or classical logic, it is assumed that either *A* is true or $\neg A$ is true, but not both. For vague predicates such as bald, this "two-valued" assumption is obviously not wholly accurate. As Jones gradually loses hair, it may be difficult to say at a certain stage whether he is bald or not bald. Fuzzy logic attempts to deal with this problem by allowing us to say that Jones is bald to some degree, where these degrees run from 1 (= completely hairy) to 0 (= completely bald).

Probability theory originated from the study of games of chance, and these still afford a good illustration of some of the basic concepts of the theory. If we roll a fair die, the probability of getting 5 is 1/6. This is written $P(5) = 1/6$. A conditional probability is the probability of a result given that something else has happened. For example, the probability of 5 given that the result was odd, is no longer 1/6, but 1/3; while the probability of 5 given that the result was even, is no longer 1/6, but 0. A conditional probability is written $P(A \mid B)$. So we have $P(5 \mid \text{odd}) = 1/3$, and $P(5 \mid \text{even}) = 0$. A related concept is independence. Two events *A* and *B* are said to be independent if the conditional probability of *A* given *B* is the same as the probability of *A*, or, in symbols, if $P(A \mid B) = P(A)$. Successive rolls of a die are normally assumed to be independent, that is to say, the probability of getting a 5 is always the same, namely 1/6, regardless of what results have appeared so far. An important concept for probability in AI is conditional independence. *A* and *B* are said to be conditionally independent given *C*, if $P(A \mid B \ \& \ C) = P(A \mid C)$.

Probability theory of course is not just applied to games of chance. Another typical problem is sampling from a particular population. Suppose we select a man at random from a population of men, and measure his height. If we write the result as *X*, *X* is said to be a random variable, because the value of *X* varies randomly from one individual to another. These values are, however, distributed in a particular fashion. So *X* is said to have a probability distribution. In the example of male heights, this will be the familiar bell-shaped curve, or normal distribution. A set of random variables is said to have a joint probability distribution.

Probabilities satisfy a standard set of axioms, known as the Kolmogorov axioms, after the mathematician Kolmogorov who first produced the formulation of these axioms which has come to be generally accepted by mathematicians. A number of basic results follow from these axioms, for example $P(A) + P(\neg A) = 1$. One of the most famous theorems to follow from the axioms is Bayes Theorem. We will consider a particular case of this theorem dealing with a hypothesis H, and some evidence E. Bayes Theorem states that

$$P(H \mid E) = P(E \mid H)P(H)/P(E)$$

The components of this formula have the following names:

$P(H)$ is called the prior or a priori probability of H
$P(E)$ is called the prior or a priori probability of E
$P(E \mid H)$ is called the likelihood
$P(H \mid E)$ is called the posterior or a posteriori probability of H.

The use of the formula to go from $P(H)$ to $P(H \mid E)$ is known as Bayesian conditionalization, or Bayesian reasoning or Bayesian updating. Bayesianism is, roughly speaking, the view that the problem of relating hypotheses to evidence can be solved by Bayesian reasoning.

There are a number of different interpretations of probability, and these can be classified as subjective and objective. An objective probability is one which is supposed to be a feature of the objective world, such as mass or electrical charge. A well-known objective interpretation of probability is the frequency interpretation. For example, to say that the probability of 5 is $1/6$ on this interpretation is taken to mean that, in a long series of rolls of the die, the result 5 will appear with a frequency of approximately $1/6$. Those who adopt this interpretation estimate their probabilities from frequency data.

A subjective probability, by contrast, is taken to be the measure of the degree of belief of a particular individual that some event will occur. For example, if I say that my subjective probability that it will rain in London tomorrow is $2/3$, this means that I believe to degree $2/3$ that it will rain in London tomorrow. A woman's degree of belief can be measured by the rate at which she is prepared to bet, or her betting quotient. It can be shown that, starting from this way of measuring belief, the standard axioms of probability can be derived.

An application of the subjective theory of probability to Bayesianism produces what is known as subjective Bayesianism. Here $P(H)$ is taken to represent the prior degree of belief of Mr. R, say, that H is true, while $P(H \mid E)$ represents his posterior degree of belief in H after he has come to know evidence E. A rational man on this approach changes his degree of belief in the light of new evidence E from $P(H)$ to $P(H \mid E)$, where the value of $P(H \mid E)$ is calculated using Bayes Theorem. These basic concepts of probability have been treated here very briefly, and a much fuller account is to be found in Gillies 2000.

Turning finally to causality, the traditional view of causality, to be found in e.g. Kant, was that, if A causes B, then B follows necessarily from A. An example would be: decapitation causes death. In the twentieth century, however, a weaker notion of causality has developed which could be called indeterminate causality. A familiar example would be: smoking causes lung cancer. This is held to be true even though many people smoke all their lives and never contract lung cancer. The sense of causality here is something like: smoking is an important factor in producing lung cancer; or perhaps: smokers have a higher probability of contracting lung cancer than nonsmokers. This notion of indeterminate causality naturally raises the question of how causality is related to probability.

References

Adams, J. B. 1976. "A probability model of medical reasoning and the MYCIN model." *Mathematical Biosciences* 32: 177–86.

Corfield, D. and Williamson, J., eds. 2001. *Foundations of Bayesianism.* Dordrecht: Kluwer.

de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., and Horrocks, J. C. 1972. "Computer-aided diagnosis of acute abdominal pain." *British Medical Journal* 2: 9–13.

Duda, R. O., Hart, P. E., and Nilsson, N. J. 1976. "Subjective Bayesian methods for rule-based inference systems." *Proceedings of the National Computer Conference* (AFIPS) 45: 1075–82.

Gaschnig, J. 1982. "Prospector: an expert system for mineral exploration." In D. Michie, ed., 1982: 47–64.

Gillies, D. A. 1996. *Artificial Intelligence and Scientific Method*. Oxford: Oxford University Press.

——. 1998a. "Confirmation theory." In D. M. Gabbay and P. Smets, eds., *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 1. Dordrecht: Kluwer, pp. 135–67.

——. 1998b. "Debates on Bayesianism and the theory of Bayesian networks." *Theoria* 64: 1–22.

——. 2000. *Philosophical Theories of Probability*. London and New York: Routledge.

Heckerman, D. 1986. "Probabilistic interpretations for MYCIN's certainty factors." In L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence*. Amsterdam: North-Holland, pp. 167–96.

Howard, R. A. and Matheson, J. E. 1984. "Influence diagrams." In R. A. Howard and J. E. Matheson, eds., *The Principles and Applications of Decision Analysis*, vol. 2. Menlo Park, CA: Strategic Decisions Group, pp. 721–62.

Howson, C. and Urbach, P. 1989. *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Jackson, P. 1986. *Introduction to Expert Systems*. Wokingham, UK: Addison-Wesley.

Kim, J. H. and Pearl, J. 1983. "A computational model for combined causal and diagnostic reasoning in inference systems." In *Proceedings of the 8th International Joint Conference on AI* (IJCAI-85), pp. 190–3.

Lauritzen, S. L. and Spiegelhalter, D. J. 1988. "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)." *Journal of the Royal Statistical Society B* 50: 157–224.

Leaper, D. J., Horrocks, J. C., Staniland, J. R., and de Dombal, F. T. 1972. "Computer-assisted diagnosis of abdominal pain using 'estimates' provided by clinicians." *British Medical Journal* 4: 350–4.

Michie, D., ed. 1982. *Introductory Readings in Expert Systems*. New York: Gordon and Breach.

Miller, D. 1994. *Critical Rationalism*. Chicago and La Salle, IL: Open Court.

Neapolitan, R. E. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. New York: John Wiley.

Ng, K. and Abramson, B. 1990. "Uncertainty management in expert systems." *IEEE Expert* 5: 29–48.

Pearl, J. 1982. "Reverend Bayes on inference engines: a distributed hierarchical approach." *Proceedings of the National conference on AI* (ASSI-82), pp. 133–6.

——. 1985a. "How to do with probabilities what people say you can't." *Proceedings of the Second IEEE Conference on AI Applications*, Miami, FL, pp. 6–12.

——. 1985b. "Bayesian networks: a model of self-activated memory for evidential reasoning." *Proceedings of the Cognitive Science Society*. New York: Ablex (Elsevier), pp. 329–34.

——. 1986. "Fusion, propagation and structuring in belief networks." *Artificial Intelligence* 29: 241–88.

——. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

——. 1993. "Belief networks revisited." *Artificial Intelligence* 59: 49–56.

——. 2001. "Bayesianism and causality, or, why I am only a half-Bayesian." In Corfield & Williamson, eds., 2001.

Popper, K. R. 1972 [1934]. *The Logic of Scientific Discovery*. London: Hutchinson.

Reichenbach, H. 1956. *The Direction of Time*. Berkeley: University of California Press.

Salmon, W. C. 1998. *Causality and Explanation*. New York, Oxford: Oxford University Press.

Shortliffe, E. H. and Buchanan, B. G. 1975. "A model of inexact reasoning in medicine." *Mathematical Biosciences* 23: 351–79.

Sucar, L. E. 1991. *Probabilistic Reasoning in Knowledge-based Vision Systems*. Ph.D. dissertation, Imperial College, University of London.

——, Gillies, D. F., and Gillies, D. A. 1993. "Objective probabilities in expert systems." *Artificial Intelligence* 61: 187–208.

Williams, P. 2001. "Probabilistic learning models." In Corfield & Williamson, eds., 2001.

Chapter 22

Game Theory: Nash Equilibrium

Cristina Bicchieri

Game theory aims to understand situations in which decision-makers interact. Chess is an example, as are firms competing for business, politicians competing for votes, jury members deciding on a verdict, animals fighting over prey, bidders competing in auctions, threats and punishments in long-term relationships, and so on. What all these situations have in common is that the outcome of the interaction depends on what the parties jointly do. Decision-makers may be people, organizations, animals, robots, or even genes. Game theory is a vast field, and some of its branches have seen a rapid development in the last few years, especially evolutionary game theory and experimental game theory, two areas where some of the most exciting research is being developed. For reasons of space, this chapter is limited to an assessment of the fundamental concept of noncooperative game theory, that of a Nash equilibrium. If we take Nash equilibrium to be a predictive tool, we run into problems both at the normative and descriptive levels. Many theorists have expressed misgivings about players' ability to infer an equilibrium from rationality principles alone, as well as their willingness to play equilibrium strategies in real life. Evolutionary game theory and experimental game theory have tried to respond to the normative and descriptive challenges, respectively. Though in what follows I shall focus solely on the normative challenges, it is to be hoped that the curious reader will find them interesting enough to feel stimulated to go much further.

Strategic Interaction

In a strategic interaction, the outcome of an action depends, among other things, upon the actions of other agents. Other agents have plans, preferences, and beliefs and, unless one is certain of which action will be chosen by another agent, one will have to form beliefs about other agents' possible choices, and even beliefs about the expectations that may guide another agent in choosing a particular action. Whereas rational choice is relatively straightforward in individual decision-making, it becomes more complicated in a strategic decision context.

A game is just the abstract, formal description of a strategic interaction. Any strategic interaction involves two or more decision-makers (players), each with two or more ways of acting (strategies), such that the outcome depends on the strategy choices of all the players. Each player has well-defined preferences among all the possible outcomes, enabling corresponding von Neumann–Morgenstern utilities (payoffs) to be assigned.[1] A game makes explicit the rules governing players' interaction, the players' feasible strategies, and their preferences over outcomes.

A possible representation of a game is in *normal form*. A normal-form game is completely defined by three elements: a list of players $i = 1, \ldots, n$; for each player i, a finite set of pure strategies S_i; a payoff function u_i that gives player i's payoff $u_i(s)$ for each n-tuple of strategies $(s_1, \ldots, s_n)$, where

$$u_i: \overset{n}{\underset{j=1}{X}} S_j \rightarrow R\,.$$

All players other than some given player i are customarily denoted as "$-i$". A player may choose to play a pure strategy, or instead he may choose to randomize over his pure strategies; a probability distribution over pure strategies is called a *mixed strategy* and is denoted by σ_i. The pure strategies over which a player randomizes are called the *support* of the resulting mixed strategy. Each player's randomization is assumed to be statistically independent of that of his opponents, and the payoffs to a mixed strategy are the expected values of the corresponding pure strategy payoffs.

The 2×2 matrix in figure 22.1 depicts a two-player normal-form game: each player picks a strategy independently, and the outcome, represented in terms of players' payoffs, is the joint product of these two strategies. The game in figure 22.1 is one of *complete information*, in that the players are assumed to know the rules of the game (which include players' strategies) and other players' payoffs. If players are allowed to enter into binding agreements before the game is played, we say that the game is *cooperative*. *Noncooperative games* instead make no allowance for the existence of an enforcement mechanism that would make the terms of the agreement binding on the players.

		2 C	2 D
1	C	3, 3	0, 4
1	D	4, 0	1, 1

Figure 22.1: A two-player normal-form game

Nash Equilibrium

Expected utility maximization has always been a building-block of game theory, but for many decades game theorists have paid little attention to the link between rational choice and strategic interaction, or how the outcome of strategic interaction can be derived from rational choices. In a well-known passage of their book, *Theory of Games and Economic Behavior*, von Neumann and Morgenstern said that rational players who know (i) all there is to know about the structure of the game they are playing, (ii) all there is to know about the beliefs and motives of the other players, (iii) that every player is rational, (iv) that every player knows (i) to (iii), (v) that every players knows (i) to (iv), and so on, will be able to infer the optimal strategy for every player. In that case, each player will behave rationally by maximizing his expected utility conditional on what he expects the others to do.

The above-quoted passage is important, because it states what could be rightly called the "central dogma" of game theory: that rational players will always jointly maximize their expected utilities, or play a *Nash equilibrium*. Nash equilibrium (Nash 1951) is the standard solution concept for noncooperative games. Informally, a Nash equilibrium specifies players' actions and beliefs such that (i) each player's action is optimal given his beliefs about other players' choices; (ii) players' beliefs are correct. Thus an outcome that is not a Nash equilibrium requires either that a player chooses a suboptimal strategy, or that some players "misperceive" the situation.

More formally, a Nash equilibrium is a vector of strategies $(\sigma_1^*, \ldots, \sigma_n^*)$, one for each of the n players in the game, such that each σ_i^* is optimal given (or is a *best reply* to) σ_{-i}^*. Note that optimality is only conditional on a fixed σ_{-i}, not on all possible σ_{-i}. A strategy that is a best reply to a given combination of the opponents' strategies may fare poorly *vis-à-vis* another strategy combination.

A common interpretation of Nash equilibrium is that of a self-enforcing agreement. Were players to agree in pre-play negotiation to play a particular strategy combination, they would have an incentive to stick to the agreement only in

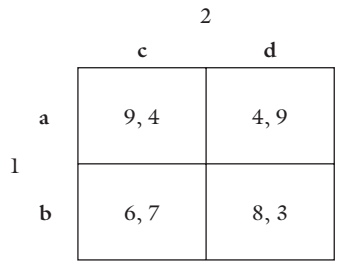

Figure 22.2: A mixed-strategy Nash equilibrium

case the agreed-upon combination is a Nash equilibrium. In the case of a *strict* Nash equilibrium, any deviation from the equilibrium strategy nets a player an inferior payoff. If the equilibrium is not strict, however, a deviation from equilibrium play may earn a player the same payoff as the equilibrium strategy. In the latter case, the incentive to follow the Nash equilibrium is less strong. The lack of a strong incentive to play one's part in a Nash equilibrium is particularly obvious in the case of mixed-strategy equilibria. Consider the game in figure 22.2.

This game has no Nash equilibrium in pure strategies but Nash proved that – provided certain restrictions are imposed on strategy sets and payoff functions – a game has at least an equilibrium in mixed strategies. Nash's result generalizes von Neumann's theorem (1928) that every game with finitely many strategies has an equilibrium in mixed strategies.

Suppose 1 plays (4/9 **a**, 5/9 **b**). Then if 2 chooses **c**, her expected utility is $4(4/9) + 7(5/9) = 17/3$. If 2 chooses **d**, she nets $9(4/9) + 3(5/9) = 17/3$. So if 1 randomizes between **a** and **b** with probabilities (4/9, 5/9), 2 is indifferent between **c**, **d**, or a lottery in which she chooses **c** with probability p and **d** with probability $(1 - p)$. Suppose 2 chooses (4/7 **c**, 3/7 **d**). In this case 1 nets 48/7 if he plays **a**, and 48/7 if he plays **b**. Hence 1 is indifferent between **a**, **b**, and any lottery (**a**p, **b**$(1 - p)$). The combination (4/9 **a**, 5/9 **b**), (4/7 **c**, 3/7 **d**) is a mixed-strategy Nash equilibrium.

In a mixed-strategy equilibrium, the equilibrium strategy of each player makes the other indifferent between the strategies on which he is randomizing. For example, if 1 were to know that 2 randomizes with probabilities (4/7, 3/7), any of his strategies (pure or mixed) would be a best reply to 2's choice and conversely, were 2 to know that 1 randomizes with probabilities (4/9, 5/9), any of her strategies, pure or mixed, would be a best reply. Paradoxically, if players agree to play a mixed-strategy equilibrium, they have no incentive to play their part in the equilibrium. A mixed-strategy equilibrium is a self-enforcing agreement only in the weak sense that – given the other players' equilibrium behavior – each player is indifferent between all the strategies (and lotteries over these strategies) in the support of her equilibrium mixed strategy.

There are, however, more serious questions raised by the Nash equilibrium concept. Ken Binmore (1987, 1988) has argued that there are two possible interpretations of Nash equilibrium. According to the *evolutive* interpretation, a Nash equilibrium is an observed regularity. Players know the equilibrium, and test the rationality of their behavior given this knowledge acquired from experience. The players (and the game theorist) can accordingly predict that a given equilibrium will be played, since they are accustomed to coordinate upon that equilibrium and expect (correctly) others to do the same. According to the more commonly adopted *eductive* interpretation instead, a game is a unique event. In this case it makes sense to ask whether players can deduce what others will do from the information available to them. The players (and the game theorist) can predict that an equilibrium will be played just in case they have enough information to infer players' choices. The standard assumptions game theorists make about players' rationality and knowledge should in principle be sufficient to guarantee that an equilibrium will obtain. The following assumptions are standard:

> CK1. The structure of the game, including players' strategy sets and payoff functions, is common knowledge among players.
> CK2. The players are rational (i.e., they are expected-utility maximizers) and this is common knowledge.

The concept of *common knowledge* was introduced by Lewis (1969), and later formalized by Aumann (1976). Simply stated, common knowledge of p among a group G means that each

		2		
		a	**b**	**c**
1	**A**	2, –1	2, 1	0, 2
	B	3, 0	1, 0	4, –1

Figure 22.3: Unique Nash equilibrium in pure strategies

member of G knows p, and each knows that each knows p, and so on *ad infinitum*. Common knowledge of rationality, preferences, and strategies may facilitate the task of predicting an opponent's strategy but, as I have argued elsewhere (Bicchieri 1993), it does not guarantee that the resulting prediction will be correct.

Consider a game that has a unique Nash equilibrium in pure strategies (figure 22.3). Can the players infer what other players will do from CK1 and CK2?

Here player 1 has two pure strategies, **A** and **B**, and player 2 has three pure strategies, **a**, **b**, and **c**. There is a unique Nash equilibrium in pure strategies, (**B**,**a**), but it is not evident that players can infer that it will be played by reasoning from CK1 and CK2. As an example of how players may reach a conclusion on how to play, consider the following argument by player 1. "If player 2 believes that I will play **A**, then it is optimal for her to pick **c**. And why would she think I play **A**? Well, she must believe that I expect her to play **b**, to which **A** is a best reply. And why would I expect her to play **b**? I would (she will think), if I were to believe she expects me to play **B** . . ." It is easy to verify that such a chain of reasoning can justify the choice of *any* strategy for both players.

The concept of Nash equilibrium embodies a notion of individual rationality, since each player's equilibrium strategy is a best reply to the opponents' strategies, but unfortunately it does not specify how players come to form the beliefs about each other's strategies that support equilibrium play. Beliefs, that is, can be internally consistent but fail to achieve the interpersonal consistency that guarantees that an equilibrium will be attained. Bernheim (1984) and Pearce (1984) have argued that assuming players' rationality (and common knowledge thereof) can only guarantee that a strategy will be *rationalizable*, in the sense of being supported by internally consistent beliefs about other players' choices and beliefs. But a combination of rationalizable strategies may not constitute a Nash equilibrium. In the game depicted in figure 22.3, all six combinations of strategies are rationalizable, yet only one of them is an equilibrium. The fact that a Nash equilibrium is always a combination of rationalizable strategies is of no help in predicting it will be played.

Note that there are strategies that are not rationalizable, in the sense of not being supported by coherent beliefs. Consider the game in figure 22.1. Suppose, again, that player 1 is deciding what to do. A quick assessment of the game will tell him that, whatever player 2 does, he is better off by choosing strategy **D**. And he must also be able to see that player 2, if rational, will never choose strategy **C**, since she, too, will always do better by playing **D**. In this case, CK1 and CK2 will lead the players to accurately predict the outcome of the game.

In a Nash equilibrium, the optimality of a strategy is only conditional on a fixed strategy combination σ_{-i}, not on all possible combinations σ_{-i}. In the game of figure 22.1 instead, the Nash equilibrium strategies (**D**, **D**) are also optimal with respect to any strategy choice of the opponent. Whatever player 1 does, player 2 is better off by choosing **D**, and the same is true of player 1. We say that a strategy s_i is *strictly dominated* by another strategy t_i if, for every choice of strategies of the other players, i's payoff from choosing t_i is strictly greater than his payoff from choosing s_i. In our example, **C** is strictly dominated by **D** for both players. A strictly dominated strategy is never rationalizable, since the belief that a player plays it is inconsistent with common knowledge of rationality. We say that s_i is *weakly dominated* by t_i if, for every choice of strategies of the other players, i's payoff from choosing t_i is at least as great as i's payoff from choosing s_i. Note that weakly dominated strategies are rationalizable, since there always exists an opponents' strategy combination to which a player's weakly dominated strategy is a best reply.

		2	
		L	**R**
1	**U**	8, 10	–100, 9
	D	7, 6	6, 5

Figure 22.4: Iterated dominance

When a game has a unique Nash equilibrium, we can predict that it will be played if we are able to show that players, armed with common knowledge of rationality and of the structure of the game, will infer the Nash solution. If players have dominated strategies, CK2 entails that they will eliminate them, and this is common knowledge (we assume that the consequences of CK1 and CK2 are common knowledge, too). Often after we have eliminated strictly dominated strategies for one player, we may find that there are now strictly dominated strategies for another player, which will be eliminated as well. This process of successive elimination can continue until there are no more strictly dominated strategies left. If a unique strategy remains for each player, we say the game has been solved by *iterated dominance*. It is easy to prove that a strategy profile thus obtained is a Nash equilibrium (Bicchieri 1993).

Consider for example the game in figure 22.4. **R** is a strictly dominated strategy for player 2, and since rationality is common knowledge, 2 is expected to eliminate **R** as a possible choice. Player 1 will now expect **L** to be played, in which case **U** dominates **D**. (**U**, **L**) is the solution to the game, and it is inferrable from CK1 and CK2. Note that assuming common knowledge of rationality (or at least some level of mutual knowledge of rationality) is crucial to obtaining the (**U**, **L**) solution. If there were some doubt about a player's rationality, the solution would unravel. For example, if 1 were to think there is a 0.01 chance that **R** is chosen, then he would be better off by choosing **D**. In real life this is likely to occur. That is, in real life a player may "play safe" and prudently choose **D**, but we are now discussing a completely different point. The question is not whether players are fully rational or believe each other to be. Rather, we want to know how far they can go in inferring a Nash solution from CK1 and CK2. As we have seen, in most cases the answer is "not far."

Predictability is hampered by another common problem encountered in game theory: Multiple Nash equilibria. Suppose two players have to divide $100 among themselves. They must restrict their proposals to integers, and each has to independently propose a way to split. If the total proposed by both is equal to or less than $100, each gets what she proposed, otherwise they get nothing. This game has 101 Nash equilibria. Is there a way to predict which one will be chosen? Alternatively, is there a way a player can infer what the other will do, and thus adjust her proposal accordingly? In real life, many people would go for the 50/50 split. It is simple, and seems equitable. In Schelling's words, it is a *focal point* (Schelling 1960). A focal point equilibrium has some property that makes it salient: A solution may be salient because of historical precedent, or because it embodies cultural norms we share (Lewis 1969). Unfortunately, mere salience is not enough to provide a player with a reason for choice. In our example, only if it is common knowledge that the 50/50 split is the salient outcome it becomes rational to propose $50. Game theory, however, filters out any social or cultural information regarding strategies, leaving players with the task of coordinating their actions on the sole basis of common knowledge of rationality (and of the structure of the game).

Consider now another game that many readers would intuitively know how to solve: figure 22.5. The game of figure 22.5 has two Nash equilibria in pure strategies: (**a**, **c**) and (**b**, **d**), but in the

		2	
		c	**d**
1	**a**	2, 2	1, 1
	b	1, 1	1, 1

Figure 22.5: Focal-point equilibrium

(**b**, **d**) equilibrium each player plays a weakly dominated strategy. (**a**, **c**) is a *Pareto-dominant* equilibrium point, since it gives both players a higher payoff than any other equilibrium in the game. For this very reason, it should be a natural focal point for both players. Should we confidently predict that (**a**, **c**) will be the solution of the game? We have seen that focal points must be common knowledge among the players before it becomes rational for them to play the focal-point equilibrium. If no such common knowledge is present, rationality alone is not a reliable guide. Could elimination of weakly dominated strategies do the trick? We know that when a player has a strictly dominated strategy, rationality dictates eliminating it, hence predicting behavior is a (relatively) simple matter. The case of weakly dominated strategies is not that straightforward. For one, a weakly dominated strategy is still a best reply to some opponent's strategy. Putting it differently, weak dominance means that there is at least one choice on the part of an opponent that makes one indifferent between the weakly dominated strategy and some other strategy. In our example, were player 2 to believe that 1 plays **b**, she would be indifferent between **c** and **d**, since both **c** and **d** are best replies to **b**; and conversely, were player 1 to expect 2 to play **d**, he would be indifferent between **a** and **b**, since both strategies are best replies to **d**.

One possible solution is to introduce a rule according to which also weakly dominated strategies should be eliminated by a rational player. Eliminating weakly dominated strategies is an example of an "eductive" procedure. When asking how the players' deductive processes might unfold, one must specify some basic principles of rationality, and then examine which choices are consistent with common knowledge of the specified principles. Such choices may or may not result in an equilibrium, but at least the link between rational choice and equilibrium (when there is such link) is made clear. The advantage of this approach is that it is possible to refine our predictions about how players might choose without assuming that they will coordinate on a particular equilibrium. Principles such as iterated strict dominance and rationalizability are examples of how it is possible to restrict the set of predictions using rationality arguments alone. In most cases, however, the set of possible outcomes is still too large.

A very different approach to the problem of indeterminacy is to start by considering the set of Nash equilibria, and ask whether some of them should be eliminated because they are in some sense "unreasonable." This is the approach taken by the *refinement* program (Kohlberg 1990, van Damme 1991).

Normal-form Refinements

Consider again the game in figure 22.5. How reasonable is the equilibrium (**b**, **d**)? Under what circumstances would players agree to play it, and then stand by the agreement? The equilibrium strategies (**b**, **d**) are weakly dominated but – as I have already argued – common knowledge of rationality does not force players to eliminate them. Prudence, however, may suggest that one should never be too sure of the opponents' choices. Even if players have agreed to play a given equilibrium, some uncertainty remains. If so, we should try to model this uncertainty in the game. Selten's insight was to treat perfect rationality as a limit case (Selten 1965). His "trembling hand" metaphor presupposes that deciding and acting are two separate processes, in that even if one decides to take a particular action, one may end up doing something else by mistake. An equilibrium strategy should be optimal not only against the opponents' strategies, but also against some very small probability $\varepsilon > 0$ that the opponents make "mistakes." Such an equilibrium is *trembling-hand perfect*. Is the equilibrium (**b**, **d**) perfect? If so, **b** must be optimal against **c** being played with probability ε and **d** being played with probability $1 - \varepsilon$ for some small $\varepsilon > 0$. But in this case the payoff to **a** is 2ε, whereas the payoff to **b** is ε. Hence for all $\varepsilon > 0$, **a** is a better strategy choice. The equilibrium (**b**, **d**) is not perfect, but (**a**, **c**) is. A prudent player therefore would discard (**b**, **d**). In this simple game, checking perfection is easy, since only one mistake is possible. With many strategies, there are many more possible mistakes to take into account. Similarly, with many players we may

need to worry about who is likely to make a mistake.

Note that the starting-point of this approach is the set of Nash equilibria of the game. It is assumed that players can calculate them, and agree to play one. The goal now is to rule out all those Nash equilibria that are not reasonable agreements. In principle, an equilibrium that is reasonable under a given criterion of reasonableness might cease to be such under another, more restrictive criterion. Specifying why an equilibrium might be unacceptable is made easier by taking into account what *would* happen if one or more players were to "deviate" from the agreed-upon solution. The reason is intuitive: a player should not agree to play his part in an equilibrium if – were the unexpected to happen – he would have been better off by playing another strategy. Therefore we may say that a crucial property required of an equilibrium is that it is *stable* to players' deviating from it. To reason about "deviations" from equilibrium it is helpful to have a richer description of the game. This is the reason why most of the refinement literature refers to games in extensive form, where the order in which players move and the information they have when making a choice are made explicit.

Games in Extensive Form

The *extensive form* of a game specifies the following information: a finite set of players $i = 1, \ldots n$, one of which might be nature (N); the order of moves; the players' choices at each move and what each player knows when she has to choose; the players' payoffs as a function of their moves; finally, moves by nature correspond to probability distributions over exogenous events. The order of play is represented by a game tree T, which is a finite set of partially ordered nodes $t \in T$ that satisfy a precedence relation denoted by "$\prec$". A *subgame* is a collection of branches of a game such that they start from the same node and the branches and the node together form a game tree by itself. In figure 22.6a, for example, player 2's decision node as well as her moves form a subgame of the original game.

Whereas normal-form games are represented by matrices, extensive-form games are represented by trees. A matrix description shows the outcomes, represented in terms of players' payoffs, for every possible combination of strategies the players might choose. A tree representation is sequential, because it shows the order in which actions are taken by the players. It is quite natural to think of sequential-move games as being ones in which players choose their strategies one after the other, and of simultaneous-move games as ones in which players choose their strategies at the same time. What is important, however, is not the temporal order of events *per se*, but whether players know about other players' actions when they have to choose their own. In the normal-form representation, players' information about other players' choices is not represented. This is the reason why a normal-form game could represent any one of several extensive-form games. When the order of play is irrelevant to a game's outcome, then restricting oneself to the normal form is justifiable. When the order of play is relevant, however, the extensive form must be specified.

In an extensive-form game, the information a player has when she is choosing an action is explicitly represented using *information sets*, which partition the nodes of the tree. If an information set contains more than one node, the player who has to make a choice at that information set will be uncertain as to which node she is at. Not knowing at which node one is means that the player does not know which action was chosen by the preceding player. If a game contains information sets that are not singletons, the game is one of *imperfect information*. It may also be the case that a player does not remember what she previously did. In this case, the game is one of *imperfect recall*. All the games we consider here, however, will be ones of perfect recall, in that players will be assumed to remember what they did and knew previously.

A *strategy* for player i is a complete plan of action that specifies an action at every node at which it is i's turn to move. Note that a strategy specifies actions even at nodes that will never be reached if that strategy is played. Consider the game in figure 22.6a. It is a finite game of perfect information in which player 1 moves first. If

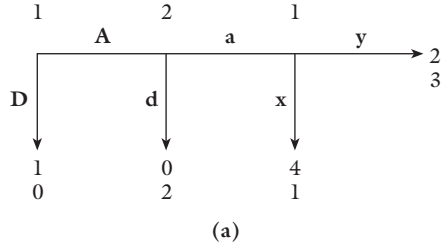

Figure 22.6a: Extensive-form game

Player 1 \ Player 2	d	a
Dx	1, 0	1, 0
Dy	1, 0	1, 0
Ax	0, 2	4, 1
Ay	0, 2	2, 3

(b)

Figure 22.6b: Normal-form game

he chooses **D** at his first node, the game ends and player 1 nets a payoff of 1, whereas player 2 gets 0. But choosing **D** at the first node is only part of a strategy for player 1. For example, it can be part of a strategy that recommends "play **D** at your first node, and **x** at your last node". Another strategy may instead recommend playing **D** at his first node, and **y** at his last decision node. Though it may seem surprising that a strategy specifies actions even at nodes that will not be reached if that strategy is played, we must remember that a strategy is a full *contingent* plan of action. For example, the strategy **Dx** recommends playing **D** at the first node, thus effectively ending the game. It is important, however, to be able to have a plan of action in case **D** is not played. Player 1 may, after all, make a mistake and, because of 2's response, find himself called to play at his very last node. In that case, having a plan helps. Note that a strategy cannot be changed during the course of the game. Though a player may conjecture about several scenarios of moves and countermoves before playing the game, at the end of deliberation a strategy must be chosen and followed through the game.

The game in figure 22.6a has two Nash equilibria in pure strategies: (**Dx**, **d**) and (**Dy**, **d**). This is easy to verify by looking at figure 22.6b, the normal-form representation of the game. Is there a way to solve the indeterminacy?

Representing the sequential version of the game as one of perfect information (figure 22.6a) helps to solve it. Suppose player 1 were to reach his last node. Since he is by assumption rational, he will choose **x**, which guarantees him a payoff of 4. Knowing (by assumption) that 1 is rational, player 2 – if she were to reach her decision node – would play **d**, since by playing **a** she would net a lower payoff. Finally, since (by assumption) player 1 knows that 2 is rational and that she knows that 1 is rational, he will choose **D** at his first decision node. The equilibrium (**Dy**, **d**) should therefore be ruled out, since it recommends an irrational move at the last node. In the normal form, both equilibria survive. The reason is simple: Nash equilibrium does not constrain behavior out of equilibrium. In our example, if 1 plans to choose **D** and 2 plans to choose **d**, it does not matter what player 1 would do at his last node, since that node will never be reached.

The sequential procedure we have used to conclude that only (**Dx**, **d**) is a reasonable solution is known as *backward induction* (Zermelo 1913). In finite games of perfect information with no ties in payoffs, backward induction always identifies a unique equilibrium. The premise of the backward-induction argument is that mutual rationality and the structure of the game are common knowledge among the players (CK1 and

CK2). It has been argued by Binmore (1987), Bicchieri (1989, 1993), and Reny (1992) that under certain conditions common knowledge of rationality leads to inconsistencies. For example, if player 2 were to reach her decision node, would she keep thinking that player 1 is rational? How would she explain 1's move? If 1's move is inconsistent with CK2, player 2 will be unable to predict future play; as a corollary, what constitutes an optimal choice at her node remains undefined. As a consequence of the above criticisms, the usual premises of backward-induction arguments have come to be questioned (Pettit & Sugden 1989; Basu 1990; Bonanno 1991).

Extensive-form Refinements

The goal of the refinement program, however, has not been the formalization of players' reasoning. The arguments proposed have been informal, their purpose being the elimination of implausible equilibria. In the normal form, Selten's trembling-hand perfection requires players to check how a strategy will perform were another player to take an action that has zero probability in equilibrium. In the extensive-form representation, players ask what would happen off-equilibrium, at points in the game tree that will never be reached if the equilibrium is played. In both cases, the starting-point is an equilibrium, which is checked for *stability* against possible deviations.

By its nature, the Nash equilibrium concept does not restrict action choices off the equilibrium path, because those choices do not affect the payoff of the player who moves there. For example, the equilibrium (**Dy**, **d**) in the game of figure 22.6 lets player 1 make an irrational choice at the last node, since that choice is not going to affect his payoff (which is determined by his choosing **D** at the beginning of the game). However, the strategy of a player at an off-equilibrium information set can affect what other players choose in equilibrium. Suppose the players consider agreeing to play (**Dy**, **d**). In order to choose **D**, player 1 must decide what would happen were he to play **A** instead. To decide whether **D** is a rational move, 1 has to think about player 2's choice at an off-equilibrium node. His conclusion about 2's choice will affect his own choice. But player 2's choice will depend upon how she interprets 1's off-equilibrium move. Player 1, in turn, must be able to anticipate 2's interpretation of his deviating from the equilibrium path. For example, if 2 were to interpret the deviation as a mistake, would she still play her part in the equilibrium (**Dy**, **d**), and choose **d**? If she expects **y** to be played at the last node, **a** is a best reply. But, at his last node, why would rational player 1 choose **y**? Is the agreement to play (**Dy**, **d**) reasonable?

The earliest refinement proposed to rule out implausible equilibria in extensive games of perfect information is *subgame perfection* (Selten 1965). A Nash equilibrium is subgame perfect if its component strategies – when restricted to any subgame – remain a Nash equilibrium of the subgame. The equilibrium (**Dy**, **d**) is not subgame perfect: in the subgame starting at the last node, **y** is a dominated strategy. Note that the backward induction equilibrium is always subgame perfect. Subgame perfection, however, only applies (nontrivially) to games that have proper subgames. Any Nash equilibrium of a game without proper subgames is trivially subgame perfect (since the whole game can be considered a subgame), but in this case the criterion does not help in resolving the indeterminacy. In figure 22.7, for example, both (**c**, **L**) and (**a**, **R**) are (trivially) subgame-perfect equilibria.

The game of figure 22.7 is a case in which we would still like to eliminate equilibria that require players to behave suboptimally in parts of the game that are reached with zero probability if a given equilibrium is played, but cannot be considered subgames. Kreps and Wilson's (1982b) *sequential equilibrium* is an answer to this problem. A sequential equilibrium is a combination of strategies and beliefs such that each player has a belief (a probability assessment) over the nodes at each of his information sets. At any information set x where i has to play – given player i's beliefs at x and the equilibrium strategies of the other players – i's strategy for the rest of the game must still maximize his expected payoff. As players move through the game tree, they rationally update their beliefs using Bayes's rule (for more on this, see Chapter 21, Probability

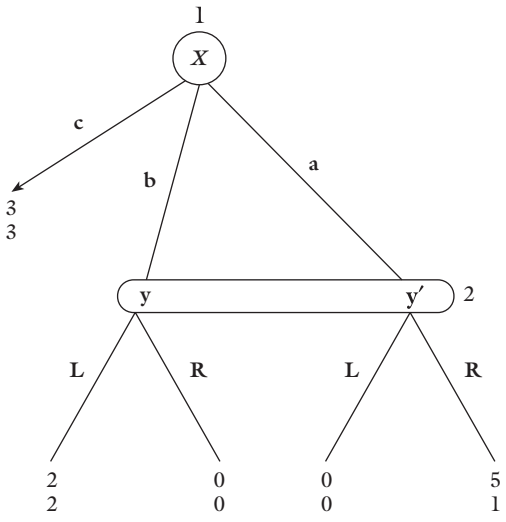

Figure 22.7: Subgame-perfect, sequential, and perfect equilibria

IN ARTIFICIAL INTELLIGENCE). The problem with the notion of sequential equilibrium is that – provided beliefs are revised according to Bayes's rule – no further restriction is imposed upon them. The consequence is that far too many Nash equilibria are still considered admissible or "reasonable." At player 2's information set, if for some reason she assigns a higher probability to node y than y', then her optimal choice is **L**. If instead she judges y and y' to be equiprobable, she will choose **R**. Thus both (**c**, **L**) and (**a**, **R**) survive as sequential equilibria, since each is supported by some acceptable belief.

Another common refinement is Selten's *perfect equilibrium* (Selten 1975). In this case players are explicitly assumed to interpret deviations from equilibrium play as "mistakes," and respond accordingly. A perfect equilibrium must be robust to small perturbations of players' equilibrium strategies. Selten's notion, however – by not imposing restrictions upon players' beliefs – lays itself open to the same criticism addressed to Kreps and Wilson's refinement. If there are several possible "mistakes" a player can make, and beliefs are unrestricted, some equilibria cannot be ruled out simply because they are supported by beliefs that make some mistakes more likely than others. Suppose that in our example player 2 believes that player 1 intends to play **c** with probability $(1 - p - q)$ very close to one, but can play **b** by mistake with higher probability ($p = 2/100$) than playing **a** by mistake ($q = 1/100$). If this is what 2 believes, she should choose **L**. Given her beliefs, **L** has an expected utility of 0.04, whereas **R** has an expected utility of 0.01. Both equilibria therefore survive some perturbations.

Note, however, that strategy **b** is strictly dominated by **c** for player 1, therefore it is highly unlikely that 1 would play **b**. This example is meant to stress the importance of restricting off-equilibrium beliefs. Such beliefs, too, should be rationally justified. The equilibrium (**c**, **L**) must be eliminated because it is supported by the belief that a dominated strategy will be played off-equilibrium, and this belief is inconsistent with common knowledge of rationality.

The need to add a condition of plausibility for off-equilibrium beliefs motivated the *forward induction* refinement (Kohlberg & Mertens 1986). Off-equilibrium beliefs, for example, should be consistent with common knowledge of rationality and any inference one may draw from it. A deviation from equilibrium should therefore be interpreted, whenever possible, as a rational move. In our example, player 1's deviation from the equilibrium (**c**, **L**) should not be interpreted as a mistake, but rather as a *signal* that he intends to play **a** (and get a higher payoff). In this case player 2 would respond with **R**. The conclusion is that equilibrium (**c**, **L**) is not robust to deviations, and should be eliminated.

As I mentioned at the outset, the refinement program attempts to establish stability criteria for Nash equilibrium. It presupposes that players will choose to play a Nash equilibrium after having eliminated several alternative equilibria on the ground that they are unreasonable. What counts as a reasonable equilibrium, however, depends upon how off-equilibrium behavior is interpreted. This, in turn, hinges on players out-of-equilibrium beliefs. So far we have developed no comprehensive theory of out-of-equilibrium behavior that indicates, for example, when a deviation should be interpreted as a signal and when as a mistake. Such theory would supply substantive (as opposed to merely formal) rationality criteria for players' beliefs, and would thus expand the traditional notion of practical rationality to

include an epistemic component. This theoretical inadequacy undermines the eductive goal of inferring (and predicting) equilibrium play from rationality principles alone.

Selection by Evolution

A Nash equilibrium need not be interpreted as a unique event. If we think of it as an observed regularity, we want to know by what process such equilibrium is reached and what accounts for its stability. When multiple equilibria are possible, we want to know why players converged to one in particular and then stayed there. An alternative way of dealing with multiple equilibria is to suppose that the selection process is made by nature.

Evolutionary theories are inspired by population biology (e.g. Maynard Smith & Price 1973). These theories dispense with the notion of the decision-maker, as well as with best responses/optimization, and use in their place a natural selection, "survival-of-the-fittest" process together with mutations to model the frequencies with which various strategies are represented in the population over time. In a typical evolutionary model, players are preprogrammed for certain strategies, and are randomly matched with other players in pair-wise repeated encounters. The relative frequency of a strategy in a population is simply the proportion of players in that population who adopt it. The theory focuses on how the strategy profiles of populations of such agents evolve over time, given that the outcomes of current games determine the frequency of different strategies in the future.

As an example, consider the symmetric game in figure 22.8 and suppose that there are only two possible behavioral types: "hawk" and "dove."[2]

A hawk always fights and escalates contests until it wins or is badly hurt. A dove sticks to displays and retreats if the opponent escalates the conflict; if it fights with another dove, they will settle the contest after a long time. Payoffs are expected changes in fitness due to the outcome of the game. Fitness here means just reproductive success (e.g., the expected number of offspring per time unit).

	H	**D**
H	$\frac{B-C}{2}, \frac{B-C}{2}$	$B, 0$
D	$0, B$	$\frac{B}{2}, \frac{B}{2}$

Figure 22.8: Hawk and dove game

Suppose injury has a payoff in terms of loss of fitness equal to C, and victory corresponds to a gain in fitness B. If hawk meets hawk, or dove meets dove, each has a 50 percent chance of victory. If a dove meets another dove, the winner gets B and the loser gets nothing, so the average increase in fitness for a dove meeting another dove is $B/2$. A dove meeting a hawk retreats, so her fitness is unchanged, whereas the hawk gets a gain in fitness B. If a hawk meets another hawk, they escalate until one wins. The winner has a fitness gain B, the loser a fitness loss C. So the average increase in fitness is $(B - C)/2$. The latter payoff is negative, since we assume the cost of injury is greater than the gain in fitness obtained by winning the contest. We assume that players will be randomly paired in repeated encounters, and in each encounter they will play the stage game of figure 22.8.

If the population were to consist predominantly of hawks, selection would favor the few doves, since hawks would meet mostly hawks and end up fighting with an average loss in fitness of $(B - C)/2$, and $0 > (B - C/2)$. In a population dominated by doves, hawks would spread, since every time they meet a dove (which would be most of the time) they would have a fitness gain of B, whereas doves on average would only get $B/2$.

Maynard Smith interpreted evolutionary games as something that goes on at the phenotypic level. The fitness of a phenotype depends on its frequency in the population. A strategy is a phenotype, and a player is just an instance of such a behavioral phenotype. In our example, we have only two behavioral phenotypes: "hawk" and "dove." Evolutionary game theory wants to know how strategies do on average when games are played repeatedly between individuals who are

randomly drawn from a large population. The average payoff to a strategy depends on the composition of the population, so a strategy may do very well (in terms of fitness) in an environment and poorly in another. If the frequency of hawks in the population is q and that of doves correspondingly $(1 - q)$, the average increase in fitness for the hawks will be $q(B - C)/2 + (1 - q)B$, and $(1 - q)B/2$ for the doves. The average payoff of a strategy in a given environment determines its future frequency in the population. Strategies that, on average, earn high payoffs in the current environment are assumed to increase in frequency, and strategies that, on average, earn lower payoffs are assumed to decrease in frequency. If the average payoffs of the different strategies are the same, then the composition of the population is stable. In our example, the average increase in fitness for the hawks will be equal to that for the doves when the frequency of hawks in the population is $q = B/C$. At that frequency, the proportion of hawks and doves is stable. If the frequency of hawks is less that B/C, then they do better than doves, and will consequently spread; if their frequency is larger than B/C, they will do worse than doves and will shrink.

Note that if $C > B$, then $(B - C)/2 < 0$, so the game in figure 22.8 has two pure-strategy Nash equilibria: (**H**, **D**) and (**D**, **H**). There is also a mixed-strategy equilibrium in which Hawk is played with probability $q = B/C$ and Dove is played with probability $(1 - q) = C - B/C$. If the game of figure 22.8 were played by rational agents who *chose* which behavior to display, we would be at a loss in predicting their choices. From CK1 and CK2 the players cannot infer that a particular equilibrium will be played; moreover, since there are no dominated strategies, all possible outcomes are rationalizable. In the hawk/dove example, however, players are not rational and do not choose their strategies. So if an equilibrium is attained it must be the outcome of some process very different from rational deliberation. The process at work is natural selection: high-performing strategies increase in frequency whereas low-performing strategies' frequency diminishes and eventually goes to zero.

We have seen that in a population composed mostly of doves, hawks will thrive, and the opposite would occur in a population composed mainly of hawks. So for example if "hawks" dominate the population, a mutant displaying "dove" behavior can invade the population, since individuals bearing the "dove" trait will do better than hawks. The main solution concept used in evolutionary game theory is the *evolutionarily stable strategy* (ESS) introduced by Maynard Smith and Price (1973). A strategy or behavioral trait is evolutionarily stable if, once it dominates in the population, it does strictly better than any mutant strategy, hence it cannot be invaded. To formalize this concept, let me first make a brief digression. In a symmetric game like hawk/dove, we have a finite set of pure strategies S and a corresponding set Δ of mixed strategies. A population state is equivalent to a mixed-strategy $x \in \Delta$. Note that the evolutionary model gives a natural interpretation to mixed strategies as the proportions of certain strategies (or traits) in a population. A state in which each individual plays a pure strategy and the proportion of different strategies correspond to x is called a polymorphic state. Alternatively, we may interpret the population state x as monomorphic, in the sense that each player plays the mixed-strategy x. In a two-player game, being matched against a randomly drawn individual in population state x is equivalent to being matched against an individual who plays the mixed strategy x. Hence the average payoff of playing strategy y in population state x is equal to the expected payoff to y when played against the mixed strategy x, i.e. $u(y, x)$. The population average in this case is equal to the expected payoff of the mixed strategy x when matched against itself, i.e. $u(x, x)$.

In a symmetric, two-player game, x is an ESS if and only if, for all $y \in \Delta$ such that $y \neq x$,

(1) $u(x, x) > u(y, x)$

or

(2) $u(x, x) = u(y, x)$, and $u(x, y) > u(y, y)$.

Condition (1) tells us that strategy x is a unique best reply against itself. If the bulk of the population consists of type x and a small number of mutants of type y enters the population, if x does better against x than y does against x, y will

be less fit and disappear. However, if x is a mixed strategy, we know (1) does not hold. In this case, for x to be an ESS, (2) must hold. If both x and y perform equally well against x, then y will be less fit than x if x does better against y that y does against y.

In the hawk/dove game, neither of the two pure behavioral types is evolutionarily stable, since each can be invaded by the other. We know, however, that a population in which there is a proportion $q = B/C$ of hawks and $(1 - q) = C - B/C$ of doves is stable. This means that the type of behavior that consists in escalating fights with probability $q = B/C$ cannot be invaded by any other type, hence it is an ESS. To show that the mixed strategy $x = (B/C, C - B/C)$ is an ESS, we have to show that condition (2) is satisfied. Indeed, $u(x, y) - u(y, y) = 1/2C(B - Cq)^2$ is greater than zero for all $q \neq B/C$.

An ESS is a strategy that, when it dominates the population, is a best reply against itself. Therefore an evolutionarily stable strategy such as $(B/C, C - B/C)$ is a Nash equilibrium. Though every ESS is a Nash equilibrium, the reverse does not hold; in our stage game, there are three Nash equilibria, but only the mixed-strategy equilibrium $(B/C, C - B/C)$ is an ESS. However, when a strategy is a *unique* best reply to itself, it is both an ESS and a *strict* Nash equilibrium. In this special case the reverse also holds: every strict Nash equilibrium is an ESS. In a strict equilibrium, there exists no other strategy which is an alternative best reply to the equilibrium strategy, and this guarantees non-invadability. As an example, consider the game in figure 22.5. There are two pure-strategy Nash equilibria, (**a**, **c**) and (**b**, **d**), but only (**a**, **c**) is strict. It is easy to verify that (**a**, **c**) consists of ESS satisfying condition (1).

The prior examples show how evolution can at least partially solve the problem of equilibrium selection without imposing heroic cognitive requirements on players. An ESS is, in fact, not just a Nash equilibrium but also a perfect and proper equilibrium (van Damme 1987). Furthermore, an evolutionary account of how a Nash equilibrium is achieved provides an explanation of the dynamics of the selection process, something which the refinement program cannot do.

In the hawk/dove example, we have assumed that the success of a strategy depends on the outcome of pairwise random matches. It is often the case that a strategy's success depends not on the strategy played by a particular opponent, but on the population-wide frequencies of strategies. When examining behavior in a *population game*, we adopt the concept of an *evolutionarily stable state* (also ESS) (Hofbauer & Sigmund 1998).

Suppose the game has N pure strategies, with an $N \times N$ symmetric expected payoff matrix $A = (a_{ij})$. There is an infinite number of players, and each player initially commits to playing exactly one of the N pure strategies. Let p be the $N \times 1$vector denoting the population-wide proportion of each of the N strategies (player types) in the population at a given time. Let

$$f_i(p) = \sum_j a_{ij} p_j = A_i p$$

denote the fitness of strategy i and let

$$\sum_i f_i(p) = Ap$$

denote the population-wide payoff. The population-wide weighted average fitness value is $p^T Ap$. We say that $\hat{p}$ is an *evolutionarily stable state* if for any $p \neq \hat{p}$ in the neighborhood of $\hat{p}$, we have:

$$\hat{p}^T Ap > p^T Ap$$

This captures the idea that the population-wide payoff under $\hat{p}$ is higher (locally) than for any other vector p.[3]

The definitions of evolutionarily stable strategies or states are static. To describe the dynamic process that leads to a certain distribution of strategies in a population, we have models of the selection dynamics that express the growth rate of a strategy i in population state p as a function of i's average payoff in p relative to the average payoff to other strategies in p. ESS do not refer to a specific dynamic, but biologists and evolutionary game theorists frequently use deterministic *replicator* dynamics (Taylor & Jonker 1978) of the form:

$$(*) \quad p_i(t+1) = \frac{p_i(t)A_ip(t)}{p^T(t)Ap(t)},$$

where $p(t)$ denotes the population-wide proportions at time t, the denominator is a measure of average strategy fitness in the population at t, and the numerator measures the fitness of strategy i at time t. Strategies with above-average fitness see their proportions increase, and those with below average fitness see their proportions decrease. Note that (*) is a deterministic system which allows some strategies to become extinct, in the sense that $p_i(t) = 0$ for some i, t. To prevent extinction, mutations are added, but a discussion of how to modify (*) to include mutations and how to interpret the latter would take us too far from the present topic. For an analysis of stochastic models, see Foster and Young (1990).

ESS are asymptotically stable fixed points of this replicator dynamic, though the converse need not be true (see e.g. Samuelson 1997). A similar relationship holds between the replicator dynamic and Nash equilibria: if $\hat{p}$ is a Nash equilibrium of the symmetric $N \times N$ game with expected payoff matrix A, then $\hat{p}$ is a stationary state of the replicator dynamic.

In evolutionary theory replication, variation and heredity are the basic assumptions. Any entity capable of replicating itself with differential success will be subject to an evolutionary process. Differential success, in turn, is related to hereditary variations. In biology, replicators are genes and in genetic evolution, variation is provided by random mutations and recombinations of gene sequences. Behavioral patterns can be replicators, too, in the sense that behavioral trait x is replicated when a gene **x** that predisposes its carriers to behave according to this pattern replicates itself. This means that bearers of gene **x** will behave in ways that make them reproductively successful, so that in the next generation there will be more copies of **x**. To the extent that behavior x promotes the replication of its predisposing gene, we are correct in saying that the behavior is replicating itself. Individuals are just bearers of such genetic material, hence they are born with fixed behavioral traits. Variation of competing strategies is provided by random mutations and recombinations of gene sequences.

When we think of strategies, however, we usually refer to behaviors that are not genetically inherited. In economic and political applications of game theory, actors can be firms, political parties, nations. Even when actors are individuals, their strategies have a strong cultural component. Evolutionary models can still be applied to explain how Nash equilibria are attained and whether they are stable, but selection mechanisms in this case work through processes of cultural transmission such as learning and imitation. Learning and imitation are subject to mistakes, and new strategies may enter the population either by random mistake or by purposeful innovation. Payoffs in this case cannot represent fitness changes, but if we give them a utility interpretation, we must provide for interpersonal comparisons of utilities. Indeed, to imitate a more successful individual, one must be able to compare one's payoffs with the payoffs of others, but traditional von Neumann–Morgenstern utilities do not allow for such comparisons.

Evolutionary games provide us with a way of explaining how agents that may or may not be rational and – if so – subject to severe information and calculation restrictions, achieve and sustain a Nash equilibrium. When there exist evolutionarily stable strategies (or states), we know which equilibrium will obtain, without the need to postulate refinements in the way players interpret off-equilibrium moves. Yet we need to know much more about processes of cultural transmission, and to develop adequate ways to represent payoffs, so that the promise of evolutionary games is actually fulfilled.

Notes

1 In a game, a player's action may have one of several possible consequences, depending on the other players' choices. It is usually assumed that the probabilities with which the consequences occur are objective and known to the decision-maker. Suppose action a has two possible consequences, x and x', which occur with probability p and $(1 - p)$, respectively. Choosing action a is like choosing a lottery that gives prize x with probability p, and prize x' with probability $(1 - p)$. Agents are assumed to have

preferences over such lotteries. If preferences are complete, transitive, and satisfy a number of other conditions (von Neumann & Morgenstern 1944), they can be represented by the expectation of a real-valued utility function $U: C \to \boldsymbol{R}$ (unique up to a positive linear transformation) such that, for any two lotteries a and b, $a \succ b$ iff $\Sigma_{x \in a} p(x) U(x) > \Sigma_{y \in b} p(y) U(y)$. A rational agent will choose an action (lottery) a^* that maximizes the expected value of a von Neumann–Morgenstern utility function.

2 A two-player game is symmetric if (a) S_1 and S_2 have the same, finite number of elements, and (b) the payoff matrix is symmetric, i.e. for all i and $j \in S$, $u_1(i, j) = u_2(j, i)$.

3 By contrast, $\hat{p}$ is a symmetric *Nash equilibrium* if $\hat{p}^T A \hat{p} \geq p^T A \hat{p}$ for all feasible p.

References

(*Note*: references marked with one asterisk are readable by an undergraduate with some basic mathematical knowledge. Two asterisks mean the reader must have greater mathematical sophistication and/or some background knowledge in game theory.)

**Aumann, R. 1976. "Agreeing to disagree." *Annals of Statistics* 4: 1236–9.

*Basu, K. 1990. "On the non-existence of a rationality definition for extensive games." *International Journal of Game Theory* 19: 33–44.

**Bernheim, B. D. 1984. "Rationalizable strategic behavior." *Econometrica* 52: 1007–28.

*Bicchieri, C. 1988. "Common knowledge and backward induction: A solution to the paradox." In M. Vardi, ed., *Theoretical Aspects of Reasoning about Knowledge*. Los Altos, CA: Morgan Kaufmann Publishers.

*——. 1989. "Self-refuting theories of strategic interaction: A paradox of common knowledge." *Erkenntnis* 30: 69–85.

*——. 1993. *Rationality and Coordination*. Cambridge: Cambridge University Press.

*Binmore, K. 1987. "Modeling rational players I." *Economics and Philosophy* 3, 179–214.

*——. 1988. "Modeling rational players II." *Economics and Philosophy* 4: 9–55.

*Bonanno, G. 1991. "The Logic of Rational Play in Games of Perfect Information." *Economics and Philosophy* 7: 37–61.

**Foster, D. and Young, P. 1990. "Stochastic evolutionary game dynamics." *Theor. Pop. Biol.* 38: 219–32.

*Harsanyi, J. and Selten, R. 1987. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

**Hofbauer, J. and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

*Kohlberg, E. 1990. "Refinement of Nash equilibrium: the main ideas." In T. Ichiishi, A. Neyman, and Y. Tauman, eds., *Game Theory and Applications*. San Diego: Academic Press.

**—— and Mertens, J.-F. 1986. "On the strategic stability of equilibria." *Econometrica* 54: 1003–37.

**Kreps, D. and Wilson, R. 1982. "Sequential Equilibria." *Econometrica* 50: 863–94.

*Lewis, D. 1969. *Convention*. Cambridge: Cambridge University Press.

*Maynard Smith, J. 1982. *Evolution and the Theory of Ga*mes. Cambridge: Cambridge University Press.

*—— and Price, G. 1973. "The logic of animal conflicts." *Nature* 246: 15–18.

**Myerson, R. 1978. "Refinements of the Nash equilibrium concept." *International Journal of Game Theory* 7: 73–80.

*Nash, J. 1951. "Noncooperative games." *Annals of Mathematics* 54: 289–95.

**Pearce, D. G. 1984. "Rationalizable strategic behavior and the problem of perfection." *Econometrica* 52: 1029–50.

*Pettit, P. and Sugden, R. 1989. "The backward induction paradox." *The Journal of Philosophy* 4: 1–14.

*Reny, P. 1992. "Rationality in extensive form games." *Journal of Economic Perspectives* 6: 92–100.

**Samuelson, L. 1997. *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press.

*Schelling, T. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

**Selten, R. 1965. "Spieltheorethische behandlung eines oligopolmodels mit nachfragetragheit." *Zietschrift fur die gesante Staatwissenschaft* 121: 301–24.

**——. 1975. "Re-examination of the perfectness concept for equilibrium points in extensive games." *International Journal of Game Theory* 4: 22–55.

**Taylor, P. D. and Jonker, L. 1978. "Evolutionary stable strategies and game dynamics." *Math. Biosci.* 40: 145–56.

**van Damme, E. 1987. *Stability and Perfection of Nash Equilibria*. Berlin: Springer Verlag.

*von Neumann, J. 1928. "Zur theorie der gesellschaftspiele." *Mathematische Annalen* 100: 295–320.

*—— and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

*Zermelo, E. 1913. "Uber eine anwendung der megenlehre auf die theorie des schachspiels." In *Proceedings of the Fifth International Congress of Mathematicians*, vol. 2. Cambridge: Cambridge University Press.

Part VII

Science and Technology

Chapter 23

Computing in the Philosophy of Science

Paul Thagard

1 Introduction

What do philosophers do? Twenty or thirty years ago, one might have been told "analyze concepts" or "evaluate arguments." The answer "write computer programs" would have inspired a blank stare and computational philosophy of science might have sounded like the most self-contradictory enterprise in philosophy since business ethics (Thagard 1988). But computer use has since become much more common in philosophy, and computational modeling can be seen as a useful addition to philosophical method, not as the abandonment of it (see Chapter 26, Computational Modeling as a Philosophical Methodology).

If philosophy consisted primarily of conceptual analysis, or mental self-examination, or investigation of *a priori* truths, then computer modeling would indeed be alien to the enterprise. However, one may argue in favor of a different picture of philosophy, as primarily concerned with producing and evaluating *theories*, for example theories of knowledge (epistemology), reality (metaphysics), and right and wrong (ethics). One of the main functions of a theory of knowledge is to explain how knowledge grows. This requires describing both the structure of knowledge and the inferential procedures by which knowledge can be increased. Although epistemologists often focus on mundane knowledge, the most impressive knowledge gained by human beings comes through the operation of science: experimentation, systematic observation, and theorizing concerning the experimental and observational results. In attempting to understand the structure and development of scientific knowledge, philosophers of science have traditionally employed a number of approaches, such as logical analysis and historical case studies. Computational modeling provides an additional method that has already advanced understanding of such traditional problems as theory evaluation and scientific discovery.

This chapter concerns how computational models are making substantial contributions to the philosophy of science. It reviews the progress made by three distinct computational approaches: cognitive modeling, engineering artificial intelligence, and theory of computation. The aim of cognitive modeling is to simulate aspects of human thinking; for philosophy of science, this becomes the aim to simulate the thinking that scientists use in the construction and evaluation of hypotheses. Much artificial intelligence research, however, is not concerned with modeling human thinking, but with constructing algorithms that

perform well on difficult tasks independently of whether the algorithms imitate human thinking. Similarly, the engineering AI approach to philosophy of science seeks to develop computational models of discovery and evaluation independently of questions of human psychology. Computational philosophy of science has thus developed two approaches that reflect the two trends in AI research, one concerned with modeling human performance and the other with machine intelligence. A third approach uses abstract mathematical analysis and applies the theory of computation to problems in the philosophy of science.

2 Cognitive Modeling

Cognitive science is the interdisciplinary study of mind, embracing philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology. From its modern origins in the 1950s, cognitive science has primarily worked with the computational-representational understanding of mind: we can understand human thinking by postulating mental representations akin to computational data structures and mental procedures akin to algorithms (Thagard 1996). The cognitive-modeling approach in computational philosophy of science views topics such as discovery and evaluation as open to investigation using the same techniques employed in cognitive science. To understand how scientists discover and evaluate hypotheses, we can develop computer models that employ data structures and algorithms intended to be analogous to human mental representations and procedures. This approach can be viewed as part of naturalistic epistemology, which sees the study of knowledge as closely tied to human psychology, not as an abstract logical exercise.

2.1 Discovery

In the 1960s and 1970s, philosophers of science discussed whether there is a "logic of discovery" and whether discovery (as opposed to evaluation) is a legitimate topic of philosophical (as opposed to psychological) investigation. In the 1980s, these debates were superseded by computational research on discovery that showed how actual cases of scientific discovery can be modeled algorithmically. Although the models that have been produced to date clearly fall well short of simulating all the thought processes of creative scientists, they provide substantial insights into how scientific thinking can be viewed computationally.

Because of the enormous number of possible solutions for any scientific problem, the algorithms involved in scientific discovery cannot guarantee that optimal discoveries will be made from input provided. Instead, computer models of discovery employ *heuristics*, approximate methods for attempting to cut through data complexity and find patterns. The pioneering step in this direction was the BACON project of Pat Langley, Herbert Simon, and their colleagues (Langley et al. 1987). BACON is a program that uses heuristics to discover mathematical laws from quantitative data, for example discovering Kepler's third law of planetary motion. Although BACON has been criticized for assuming an oversimple account of human thinking, Qin and Simon (1990) found that human subjects could generate laws from numerical data in ways quite similar to BACON.

Scientific discovery produces qualitative as well as quantitative laws. Kulkarni and Simon (1988) produced a computational model of Krebs' discovery of the urea cycle. Their program, KEKADA, reacts to anomalies, formulates explanations, and carries out simulated experiments in much the way described in Hans Krebs' laboratory notebooks.

Not all scientific discoveries are as data-driven as the ones discussed so far. They often involve the generation of new concepts and hypotheses that are intended to refer to non-observable entities. Thagard (1988) developed computational models of conceptual combination, in which new theoretical concepts such as *sound wave* are generated, and of abduction, in which new hypotheses are generated to explain puzzling phenomena. Magnani (2001) has also discussed how abductive inference can produce discoveries in science and mathematics. Darden (1990, 1998) has investigated computationally how theories that have empirical problems can be repaired.

One of the most important cognitive mechanisms for discovery is analogy, since scientists often make discoveries by adapting previous knowledge to a new problem. Analogy played a role in some of the most important discoveries ever made, such as Darwin's theory of evolution and Maxwell's theory of electromagnetism. During the 1980s, the study of analogy went well beyond previous philosophical accounts through the development of powerful computational models of how analogs are retrieved from memory and mapped to current problems to provide solutions. Falkenhainer, Forbus, and Gentner (1989) produced SME, the Structure Mapping Engine, and this program was used to model analogical explanations of evaporation and osmosis (Falkenhainer 1990). Holyoak and Thagard (1989) used different computational methods to produce ACME, the Analogical Constraint Mapping Engine, which was generalized into a theory of analogical thinking that applies to scientific as well as everyday thinking (Holyoak & Thagard 1995).

The above research projects illustrate how thought processes, such as those involved in numerical law generation, theoretical concept formation, abduction, and analogy, can be understood computationally. Examples of nonpsychological investigations of scientific discovery are described in the sections on engineering AI and theory of computation.

2.2 Evaluation

How scientific hypotheses are evaluated has been a central problem in philosophy of science since the nineteenth-century debates between John Stuart Mill and William Whewell. Work in the logical positivist tradition has centered on the concept of confirmation, asking what it is for hypotheses to be confirmed by observations. More recently, various philosophers of science have taken a Bayesian approach to hypothesis evaluation, using probability theory to analyze scientific reasoning. In contrast, it is possible to develop an approach to hypothesis evaluation that combines philosophical ideas about explanatory coherence with a connectionist (neural network) computational model (Thagard 1992, 2000).

Coherence theories of knowledge, ethics, and even truth have been popular among philosophers, but the notion of coherence is usually left rather vague. Hence coherence theories do not appear sufficiently rigorous when compared to theories couched more formally using deductive logic or probability theory. But connectionist models show how coherence ideas can be precisely and efficiently implemented. Since the mid-1980s, connectionist (neural network, PDP) models have been very influential in cognitive science. Only loosely analogous to the operation of the brain, such models have numerous units that are roughly like neurons, connected to each other by excitatory and inhibitory links of varying strengths. Each unit has an activation value that is affected by the activations of the units to which it is linked, and learning algorithms are available for adjusting the strengths on links in response to experience.

In ECHO, a connectionist computational model of explanatory coherence developed in Thagard (1992), units are used to represent propositions that can be hypotheses or descriptions of evidence, and links between units to represent coherence relations. For example, if a hypothesis explains a piece of evidence, then ECHO places an excitatory link between the unit representing the hypothesis and the unit representing the evidence. If two hypotheses are contradictory or competing, then ECHO places an inhibitory link between the units representing the two hypotheses. Repeatedly adjusting the activations of the units based on their links with other units results in a resting state in which some units are on (hypotheses accepted) and other units are off (hypotheses rejected). ECHO has been used to model many important cases in the history of science (Nowak & Thagard 1992a, 1992b; Thagard 1991, 1992, 1999). Eliasmith and Thagard (1997) argued that ECHO provides a better account of hypothesis evaluation than available Bayesian accounts, and challenged proponents of Bayesian models to produce simulations of theory choice that are as detailed and historically accurate as existing ECHO simulations. ECHO has also been used in a computational model of scientific consensus in which a group of scientists reach agreement about what theory to adopt by exchanging information

about data, hypotheses, and explanations (Thagard 1999).

A different connectionist account of inference to best explanation is given by Churchland (1989). He conjectures that abductive discovery and inference to the best explanation can both be understood in terms of prototype activation in distributed connectionist models, i.e. ones in which concepts and hypotheses are not represented by individual units but by patterns of activation across multiple units. There is considerable psychological evidence that distributed representations and prototypes are important in human cognition, but no one has yet produced a running computational model of hypothesis evaluation using these ideas. Nonconnectionist models of hypothesis evaluation, including probabilistic ones, are discussed in the next section.

3 Engineering AI

The cognitive-modeling approach to computational philosophy of science allies philosophy of science with cognitive science and naturalistic epistemology, and it can be very fruitful, as the previous section shows. However, much valuable work in AI and philosophy has been done that makes no claims to psychological plausibility. One can set out to build a scientist without trying to reverse-engineer a human scientist. The engineering AI approach to computational philosophy of science is allied, not with naturalistic, psychologistic epistemology, but with what has been called "android epistemology," the epistemology of machines that may or may not be built like humans (Ford, Glymour, & Hayes 1995). This approach is particularly useful when it exploits such differences between digital computers and humans as the capacity for very fast searches to perform tasks that human scientists cannot do very well.

3.1 Discovery

One goal of engineering AI is to produce programs that can make discoveries that have eluded humans. Bruce Buchanan, who was originally trained as a philosopher before moving into AI research, reviewed over a dozen AI programs that formulate hypotheses to explain empirical data (Buchanan 1983). One of the earliest and most impressive programs was DENDRAL, which performed chemical analysis. Given spectroscopic data from an unknown organic chemical sample, it determined the molecular structure of the sample (Lindsay et al. 1980). The program META-DENDRAL pushed the discovery task one step farther back: given a collection of analytic data from a mass spectrometer, it discovered rules explaining the fragmentation behavior of chemical samples. Buchanan has continued to work on computational models of hypothesis formation in science (Buchanan & Philips 2001).

Ideally, discovery programs should be capable of advancing science by producing novel results. One of the most successful in this regard is a program for chemical discovery, MECHEM, which automates the task of finding mechanism for chemical reactions. Given experimental evidence about a reaction, the program searches for the simplest mechanism consistent with theory and experiment (Valdés-Pérez 1994, 1995). Valdés-Pérez has also written programs that have contributed to discoveries in biology and physics. Kocabas and Langley (2000) have developed a computational aid for generating process explanations in nuclear astrophysics.

In order to model biologists' discoveries concerning gene regulation in bacteria, Karp (1990) wrote a pair of programs, GENSIM and HYPGENE. GENSIM was used to represent a theory of bacterial gene regulation, and HYPGENE formulates hypotheses that improve the predictive power of GENSIM theories given experimental data. More recently, Karp has shifted from modeling historical discoveries to the attempt to write programs that make original discoveries from large scientific databases, such as ones containing information about enzymes, proteins, and metabolic pathways (Karp & Mavrovouniotis 1994). A 1997 special issue of the journal *Artificial Intelligence* (vol. 91) contains several examples of recent work on computational discovery.

Cheeseman (1990) used a program that applied Bayesian probability theory to discover

previously unsuspected fine structure in the infrared spectra of stars. Machine learning techniques are also relevant to social science research, particularly the problem of inferring causal models from social data. The TETRAD program looks at statistical data in fields such as industrial development and voting behavior and builds causal models in the form of a directed graph of hypothetical causal relationships (Glymour, Scheines, Spirtes, & Kelly 1987; Spirtes, Glymour, & Scheines 1993).

One of the fastest-growing areas of artificial intelligence is "data mining," in which machine learning techniques are used to discover regularities in large computer data bases such as the terabytes of image data collected by astronomical surveys (Fayyad, Piatetsky-Shapiro, & Smyth 1996). Data mining is being applied with commercial success by companies that wish to learn more about their operations, and similar machine learning techniques may have applications to large scientific data bases such as those being produced by the human genome project.

3.2 Evaluation

The topic of how scientific theories can be evaluated can also be discussed from a computational perspective. Many philosophers of science (e.g. Howson & Urbach 1989) adopt a Bayesian approach to questions of hypothesis evaluation, attempting to use probability theory to describe and prescribe how scientific theories are assessed. But computational investigations of probabilistic reasoning must deal with important problems involving tractability that are usually ignored by philosophers. A full-blown probabilistic approach to a problem of scientific inference would need to establish a full joint distribution of probabilities for all propositions representing hypotheses and evidence, which would require 2^n probabilities for n hypotheses, quickly exhausting the storage and processing capacities of any computer. Ingenious methods have been developed by computer scientists to avoid this problem by using causal networks to restrict the number of probabilities required and to simplify the processing involved (Pearl 1988, Neapolitain 1990). Surprisingly, such methods have not been explored by probabilistic philosophers of science, who have tended to ignore the substantial problem of the intractability of Bayesian algorithms.

Theory evaluation in the context of medical reasoning has been investigated by a group of artificial intelligence researchers at Ohio State University (Josephson & Josephson 1994). They developed a knowledge-based system called RED that uses data concerning a patient's blood sample to infer what red-cell antibodies are present in the patient. RED performs an automated version of inference to the best explanation, using heuristics to form a composite hypothesis concerning what antibodies are present in a sample. Interestingly, Johnson and Chen (1996) compared the performance of RED with the performance of the explanatory coherence program ECHO on a set of 48 cases interpreted by clinical experts. Whereas RED produced the experts' judgments in 58 percent of the cases, ECHO was successful in 73 percent of the cases. Hence, although the engineering AI approach to scientific discovery has some evident advantages over the cognitive-modeling approach in dealing with some problems, such as mining hypotheses from large data bases, the cognitive-modeling approach exemplified by ECHO has not yet been surpassed by a probabilistic or other program that ignores human performance.

4 Theory of Computation

Both the cognitive-modeling and engineering AI approaches to philosophy of science involve writing and experimenting with running computer programs. But it is also possible to take a more theoretical approach to computational issues in the philosophy of science, exploiting results in the theory of computation to reach conclusions about processes of discovery and evaluation.

4.1 Discovery

Scientific discovery can be viewed as a problem in formal-learning theory, in which the goal is to identify a language given a string of inputs (Gold 1968). Analogously, a scientist can be

thought of as a function that takes as input a sequence of formulas representing observations of the environment and produces as output a set of formulas that represent the structure of the world (Jain, Osherson, Royer, & Sharma 1999; Kelly & Glymour 1989; Martin & Osherson 1998; Osherson & Weinstein 1989). Although formal-learning theory has produced some interesting theorems, they are limited in their relevance to the philosophy of science in several respects. Formal-learning theory assumes a fixed language and therefore ignores the conceptual and terminological creativity that is important to scientific development. In addition, formal-learning theory tends to view hypotheses produced as a function of input data, rather than as a much more complex function of the data and the background concepts and theories possessed by a scientist. Formal-learning theory also overemphasizes the goal of science to produce true descriptions, neglecting the important role of explanatory theories and hypothetical entities in scientific progress.

Nevertheless, ongoing work in formal-learning theory may shed light on more realistic kinds of scientific discovery. Schulte (2000) has provided a formal account of a problem in physics that starts with observed reactions and infers conservation principles that govern all reactions among elementary particles. He shows that there is a reliable inference procedure that is guaranteed to arrive at an empirically adequate set of conservation principles as more and more evidence is obtained. Kelly (1996) also discusses more realistic methods for modeling science, which are reviewed in the next section.

4.2 Evaluation

The theory of computational complexity has provided some interesting results concerning hypothesis evaluation. Suppose you have n hypotheses and you wish to evaluate all the ways in which combinations of them can be accepted and rejected: you then have to consider 2^n possibilities, an impossibly large number for even not very large n. Bylander et al. (1991) gave a formal definition of an abduction problem consisting of a set of data to be explained and a set of hypotheses to explain them. They then showed that the problem of picking the best explanation is NP-hard, i.e. it belongs to a class of problems that are generally agreed by computational theorists to be intractable in that the amount of time to compute them increases exponentially as the problems grow in size (see Chapter 2, Complexity). Similarly, Thagard and Verbeurgt (1998) and Thagard (2000) generalized explanatory coherence into a mathematical coherence problem that is NP-hard. What these results show is that theory evaluation, whether it is conceived in terms of Bayesian probabilities, heuristic assembly of hypotheses, or explanatory coherence, must be handled by computational approximation, not through exhaustive algorithms. So far, the theoretical results concerning coherence and scientific evaluation have been largely negative, but they serve to outline the limits within which computational modeling must work.

Kelly (1996, 2001) has approached the problem of evaluation from a different epistemological and formal perspective. In contrast to the view that scientific inference adopts theories on the basis of their coherence with data and each other, he suggests that philosophy of science should aim to specify reliable procedures that are *guaranteed* to converge to correct outputs. For Kelly, it is not enough to be able to say when one theory is better than another; rather, a formal specification of scientific methods should provide algorithms that come with a guarantee that they will converge on the right answers, as in Schulte's (2000) method of inferring conservation laws. Although Kelly has proved some interesting theorems concerning the conditions under which such guarantees might be available, it is not clear whether scientific inference in general is open to this kind of analysis. Inference from numerical data to mathematical laws that describe them may use methods that can be proven to be reliable, but inference to theories that postulate such theoretical entities as quarks, genes, and mental representations seems inescapably risky. There is good empirical evidence, particularly in the enormously successful technological applications of theories in the natural sciences, that scientific method does sometimes converge on an approximation to truth. But it is unlikely that guarantees of reliability and

convergence to the truth will turn out to be available for complex, theory-oriented kinds of scientific inquiry.

5 What Computing Adds to Philosophy of Science

More than twenty years ago, Aaron Sloman (1978) published an audacious book, *The Computer Revolution in Philosophy*, which predicted that within a few years any philosopher not familiar with the main developments of artificial intelligence could fairly be accused of professional incompetence. Since then, computational ideas have had a substantial impact on the philosophy of mind, but a much smaller impact on epistemology and philosophy of science. Why? One possible reason is the kind of training that most philosophers have, which includes little preparation for actually doing computational work. Philosophers of mind have often been able to learn enough about artificial intelligence to discuss it, but for epistemology and philosophy of science it is much more useful to perform computations rather than just to talk about them. Thus this chapter can end with a summary of what is gained by adding computational modeling to the philosophical tool kit.

Bringing artificial intelligence into philosophy of science introduces new conceptual resources for dealing with the structure and growth of scientific knowledge. Instead of being restricted to the usual representational schemes based on formal logic and ordinary language, computational approaches to the structure of scientific knowledge can include many useful representations such as prototypical concepts, concept hierarchies, production rules, causal networks, mental images, dynamic models, and so on. Philosophers concerned with the growth of scientific knowledge from a computational perspective can go beyond the narrow resources of inductive logic to consider algorithms for generating numerical laws, discovering causal networks, forming concepts and hypotheses, and evaluating competing explanatory theories.

AI not only provides new conceptual resources to philosophy of science, it also brings a new methodology involving the construction and testing of computational models. This methodology typically has numerous advantages over pencil-and-paper constructions. First, it requires considerable precision, in that to produce a running program the structures and algorithms postulated as part of scientific cognition need to be explicitly and carefully specified. Second, getting a program to run provides a test of the feasibility of its assumptions about the structure and processes of scientific development. Contrary to the popular view that clever programmers can get a program to do whatever they want, producing a program that mimics aspects of scientific cognition is often very challenging, and production of a program provides a minimal test of computational feasibility. Moreover, the program can then be used for testing the underlying theoretical ideas by examining how well the program works on numerous examples of different kinds. Comparative evaluation becomes possible when different programs accomplish a task in different ways: running the programs on the same data allows evaluation of their computational models and background theoretical ideas. Third, if the program is intended as part of a cognitive model, it can be assessed in terms of how well it models human thinking.

The assessment of cognitive models can address questions such as the following:

Genuineness: is the model a genuine instantiation of the theoretical ideas about the structure and growth of scientific knowledge, and is the program a genuine implementation of the model?
Breadth of application: does the model apply to lots of different examples, not just a few that have been cooked up to make the program work?
Scaling: does the model scale up to examples that are considerably larger and more complex than the ones to which it has been applied?
Qualitative fit: does the computational model perform the same kinds of tasks that people do in approximately the same way?
Quantitative fit: can the computational model simulate quantitative aspects of psychological experiments, e.g. ease of recall and mapping in analogy problems?
Compatibility: does the computational model simulate representations and processes that are compatible with those found in theoretical

accounts and computational models of other kinds of cognition?

Computational models of the thought processes of scientists that satisfy these criteria have the potential to increase our understanding of the scientific mind enormously. Engineering AI need not address questions of *qualitative* and *quantitative fit* with the results of psychological experiments, but should employ the other four standards of assessment.

There are numerous issues connecting computation and the philosophy of science that have not been touched on in this review. Computer science can itself be a subject of philosophical investigation, and some work has been done discussing epistemological issues that arise in computer research (see e.g. Colburn 2000; Fetzer 1998; Floridi 1999; Thagard 1993). In particular, the philosophy of artificial intelligence and cognitive science are fertile areas of philosophy of science. Computer modeling can also be useful in developing fields of interest to both philosophy and science, such as artificial life (see Chapter 15, Artificial Life, and Chapter 26, Computational Modeling as a Philosophical Methodology). The focus of this chapter has been more narrow, on how computational models can contribute to philosophy of science.

By way of conclusion, here is a list of some open problems that seem amenable to computational/philosophical investigation:

In scientific discovery, how are new questions generated? Formulating a useful question such as "How might species evolve?" or "Why do the planets revolve around the sun?" is often a prerequisite to more data-driven and focused processes of scientific discovery, but no computational account of scientific question generation has yet been given.

What is the role of emotions in scientific thinking? Scientists often generate questions in part through emotional stimuli such as curiosity and surprise, but no computer simulations of scientific thinking have yet taken emotions into account. Emotions are also outputs from scientific evaluation, as when scientists praise a theory as elegant or beautiful or exciting. Thagard 2002 discusses emotions and inputs and outputs in scientific discovery and evaluation.

What role does visual imagery play in the structure and growth of scientific knowledge? Although various philosophers, historians, and psychologists have documented the importance of visual representations in scientific thought, existing computational techniques have not been well-suited for providing detailed models of the cognitive role of pictorial mental images (see e.g. Shelley 1996). Computational models of high-level visual cognition are beginning to be developed (e.g. Davies & Goel 2000; Croft & Thagard 2002), but they have not yet been applied to scientific discovery and evaluation.

Perhaps problems such as these will, like other issues concerning discovery and evaluation, yield to computational approaches that involve cognitive modeling, engineering AI, and the theory of computation.

Acknowledgments

This chapter is an updated and expanded version of Thagard 1998. I am grateful to Luciano Floridi and Kevin Kelly for comments on earlier versions, and to the Natural Sciences and Engineering Research Council of Canada for grant support.

References

Buchanan, B. 1983. "Mechanizing the search for explanatory hypotheses." *PSA 1982*, vol. 2. East Lansing: Philosophy of Science Association. [An early paper on the prospects for mechanizing discovery.]

Buchanan, B. G. and Phillips, J. P. 2001. "Toward a computational model of hypothesis formation and model building in science." Unpublished manuscript, University of Pittsburgh. [Describes recent computational work on discovery.]

Colburn, T. R. 2000. *Philosophy and Computer Science*. Armonk: M. E. Sharpe. [Introduction to philosophical issues in computer science.]

Bylander, T., Allemang, D., Tanner, M., and Josephson, J. 1991. "The computational complexity of abduction." *Artificial Intelligence*, 49: 25–60. [Proves that abductive inference is NP-hard.]

Cheeseman, P. 1990. "On finding the most probable model." In J. Shrager and P. Langley, eds.,

Computational Models of Scientific Discovery and Theory Formation. San Mateo, CA: Morgan Kaufmann, pp. 73–96. [Applies a Bayesian approach to machine learning.]

Churchland, P. 1989. *A Neurocomputational Perspective*. Cambridge, MA: MIT Press. [Takes a neural-network approach to the philosophy of science.]

Croft, D. and Thagard, P. 2002. "Dynamic imagery: a computational model of motion and visual imagery." In L. Magnani, ed., *Model-based Reasoning: Scientific, Technology, Values*. New York: Kluwer/Plenum. [Develops algorithms for modeling visual reasoning.]

Darden, L. 1990. "Diagnosing and fixing fault in theories." In J. Shrager and P. Langley, eds., *Computational Models of Discovery and Theory Formation*. San Mateo, CA: Morgan Kaufman, pp. 219–246. [Discusses how anomalies in scientific theories can be resolved.]

——. 1998. "Anomaly-driven redesign: computational philosophy of science experiments." In T. W. Bynum and J. H. Moor, eds., *The Digital Phoenix: How Computers Are Changing Philosophy*. Oxford: Blackwell, pp. 62–78. [Describes a computer model of theory revision.]

Davies, J. R. and Goel, A. K. 2000. "A computational theory of visual analogical transfer." Technical report GIT-COGSCI-2000/3, Georgia Institute of Technology. [Provides a computational account of visual analogy.]

Eliasmith, C. and Thagard, P. 1997. "Waves, particles, and explanatory coherence." *British Journal for the Philosophy of Science* 48: 1–19. [Uses the computer program ECHO to simulate reasoning in physics.]

Falkenhainer, B. 1990. "A unified approach to explanation and theory formation." In J. Shrager and P. Langley, eds., *Computational Models of Discovery and Theory Formation*. San Mateo, CA: Morgan Kaufman pp. 157–196. [Describes a rich simulation of scientific reasoning.]

——, Forbus, K. D., and Gentner, D. 1989. "The structure-mapping engine: algorithms and examples." *Artificial Intelligence* 41: 1–63. [Presents a powerful computational approach to analogy.]

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From data mining to knowledge discovery in databases." *AI Magazine* 17(3): 37–54. [A survey of applied machine learning.]

Fetzer, J. H. 1998. "Philosophy and computer science: feflections on the program verification debate." In T. W. Bynum and J. H. Moor, eds., *The Digital Phoenix: How Computers Are Changing Philosophy*. Oxford: Blackwell, pp. 253–73. [Philosophical reflection on a live issue in computer science.]

Floridi, L. 1999. *Philosophy and Computing: An Introduction*. London: Routledge. [Discusses philosophical issues about computers.]

Ford, K. M., Glymour, C., and Hayes, P. J., eds. 1995. *Android Epistemology*. Menlo Park, CA: AAAI Press. [Collection of articles on non-psychological approaches to AI.]

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. 1987. *Discovering Causal Structure*. Orlando: Academic Press. [Provides computational methods for discovering causal structure from numerical data.]

Gold, E. 1968. "Language identification in the limit." *Information and Control* 10: 447–74. [Early seminal paper on formal-learning theory.]

Holyoak, K. J. and Thagard, P. 1989. "Analogical mapping by constraint satisfaction." *Cognitive Science* 13: 295–355. [Describes a connectionist model of analogical reasoning.]

—— and Thagard, P. 1995. *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press/Bradford Books. [Surveys cognitive science research on analogy from the perspective of the authors' multiconstraint theory.]

Howson, C. and Urbach, P. 1989. *Scientific Reasoning: The Bayesian Tradition*. Lasalle, IL: Open Court. [Defends a Bayesian approach to the philosophy of science.]

Jain, S., Osherson, D., Royer, J. S., and Sharma, A. 1999. *Systems That Learn*, 2nd ed. Cambridge, MA: MIT Press. [An up-to-date treatment of formal-learning theory.]

Johnson, T. R. and Chen, M. 1995. "Comparison of symbolic and connectionist approaches for multiple disorder diagnosis: heuristic search vs. explanatory coherence." Unpublished manuscript, Ohio State University. [A computational evaluation of different approaches to diagnosis.]

Josephson, J. R. and Josephson, S. G., eds. 1994. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge: Cambridge University Press. [Presents results of an extensive research project on abduction.]

Karp, P. and Mavrovouniotis, M. 1994. "Representing, analyzing, and synthesizing biochemical pathways." *IEEE Expert* 9(2): 11–21. [Describes a computational model of biological reasoning.]

Kelly, K. 1996. *The Logic of Reliable Inquiry*. New York: Oxford University Press. [Uses formal methods to analyze how scientists can achieve true beliefs.]

——. 2001. "The logic of success." *British Journal for the Philosophy of Science*, 51: 639–66. [Uses formal techniques to analyze the success of scientific inference.]

—— and Glymour, C. 1989. "Convergence to the truth and nothing but the truth." *Philosophy of Science* 56: 185–220. [Applies formal-learning theory to the philosophy of science.]

Kocabas, S. and Langley, P. 2000. "Computer generation of process explanations in nuclear astrophysics." *International Journal of Human-Computer Studies* 53: 377–92. [Use mathematical techniques to make astrophysical discoveries.]

Kulkarni, D. and Simon, H. 1988. "The processes of scientific discovery: the strategy of experimentation." *Cognitive Science* 12: 139–75. [Simulates Krebs' chemical discoveries.]

Langley, P., Simon, H., Bradshaw, G., and Zytkow, J. 1987. *Scientific Discovery*. Cambridge, MA: MIT Press/Bradford Books. [The pioneering work on cognitive modeling of discovery.]

Lindsay, R., Buchanan, B., Feigenbaum, E., and Lederberg, J. 1980. *Applications of Organic Chemistry for Organic Chemistry: The DENDRAL Project*. New York: McGraw Hill. [Classic early AI work on scientific reasoning.]

Magnani, L. 2001. *Abduction, Reason, and Science: Processes of Discovery and Explanation*. New York: Kluwer/Plenum. [A philosophical discussion of inference to explanatory hypotheses.]

Martin, E. and Osherson, D. 1998. *Elements of Scientific Inquiry*. Cambridge, MA: MIT Press. [Basic work in formal-learning theory.]

Neapolitain, R. 1990. *Probabilistic Reasoning in Expert Systems*. New York: John Wiley. [An introduction to Bayesian networks.]

Nowak, G. and Thagard, P. 1992a. "Copernicus, Ptolemy, and explanatory coherence." In R. Giere, ed., *Cognitive Models of Science*, vol. 15. Minneapolis: University of Minnesota Press, pp. 274–309. [Simulates the acceptance of Copernicus over Ptolemy.]

—— and ——. 1992b. "Newton, Descartes, and explanatory coherence." In R. Duschl and R. Hamilton, eds., *Philosophy of Science, Cognitive Psychology and Educational Theory and Practice*. Albany: SUNY Press, pp. 69–115. [Simulates the acceptance of Newton's physics over Descartes's.]

Osherson, D. and Weinstein, S. 1989. "Identifiable collections of countable structures." *Philosophy of Science* 56: 94–105. [Applies formal-learning theory to the philosophy of science.]

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufman. [Classic work on Bayesian networks.]

Qin, Y. and Simon, H. 1990. "Laboratory replication of scientific discovery processes." *Cognitive Science* 14: 281–312. [Experimental confirmation of computational theory of discovery.]

Schulte, O. 2000. "Inferring conservation laws in particle physics: a case study in the problem of induction." *British Journal for the Philosophy of Science* 51: 771–806. [Provides a formal model of scientific induction.]

Shelley, C. P. 1996. "Visual abductive reasoning in archaeology." *Philosophy of Science* 63: 278–301. [Analyzes interesting examples of hypothesis formation using visual imagery.]

Sloman, A. 1978. *The Computer Revolution in Philosophy*. Atlantic Highlands, NJ: Humanities Press. [An early work linking philosophy with AI.]

Spirtes, P., Glymour, C., and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag. 2nd ed. 2000, published by MIT Press. [Describes formal methods for deriving causal models from data.]

Thagard, P. 1988. *Computational Philosophy of Science*. Cambridge, MA: MIT Press/Bradford Books. [Develops computational methods in the philosophy of science for both discovery and evaluation.]

——. 1991. "The dinosaur debate: explanatory coherence and the problem of competing hypotheses." In J. Pollock and R. Cummins, eds., *Philosophy and AI: Essays at the Interface*. Cambridge, MA: MIT Press/Bradford Books, pp. 279–300. [Simulates different views on why dinosaurs became extinct.]

——. 1992. *Conceptual Revolutions*. Princeton: Princeton University Press. [Provides a cognitive-computational account of scientific revolutions.]

——. 1993. "Computational tractability and conceptual coherence: why do computer scientists believe that P ≠ NP?" *Canadian Journal of Philosophy* 23: 349–64. [Discusses grounds for current beliefs about computational complexity.]

——. 1996. *Mind: Introduction to Cognitive Science*. Cambridge, MA: MIT Press. [A concise interdisciplinary textbook.]

——. 1998. "Computation and the philosophy of science." In T. W. Bynum and J. H. Moor, eds., *The Digital Phoenix: How Computers Are Changing Philosophy*. Oxford: Blackwell, pp. 48–61. [Earlier version of the present article.]

——. 1999. *How Scientists Explain Disease*. Princeton: Princeton University Press. [Discusses the development and acceptance of the bacterial theory of ulcers in cognitive/computational terms.]

——. 2000. *Coherence in Thought and Action*. Cambridge, MA: MIT Press. [Applies a computational model of coherence to scientific and other kinds of thinking.]

——. 2002. "The passionate scientist: emotion in scientific cognition." In P. Carruthers, S. Stich, and M. Siegal, eds., *The Cognitive Basis of Science*. Cambridge: Cambridge University Press. [Describes the role of emotion in scientific thinking.]

—— and Verbeurgt, K. 1998. "Coherence as constraint satisfaction." *Cognitive Science* 22: 1–24. [A mathematical/compuational analysis of coherence.]

Valdés-Pérez, R. E. 1994. "Conjecturing hidden entities via simplicity and conservation laws: Machine discovery in chemistry." *Artificial Intelligence* 65: 247–80. [Presents powerful computational techniques for making original scientific discoveries.]

——. 1995. "Machine discovery in chemistry: new results." *Artificial Intelligence* 74: 191–201. [Describes additional computational discoveries in chemistry.]

Chapter 24

Methodology of Computer Science

Timothy Colburn

Introduction

Science and philosophy are often distinguished by pointing out that science seeks explanation while philosophy seeks justification. To ask what accounts for the neuronal firing of synapses in the brain, for example, is a scientific question, while to ask what would constitute adequate grounds for believing that an artificially constructed neural network is conscious is a philosophical one. So philosophy has been characterized as the critical evaluation of beliefs through the analysis of concepts in a given area of inquiry. Of course, science is also concerned with critically evaluating beliefs and analyzing concepts. However, philosophy is a non-empirical, or *a priori*, discipline, in distinct contrast with science.

Computer science would seem to be distinguished from philosophy just as any other science. But computer science is unique among the sciences in the types of models it creates. In seeking explanations, science often constructs models to test hypotheses for explaining phenomena. These models, in the form of experimental apparatus, are of course physical objects. The models built and manipulated in computer science, however, are not physical at all. Computer science is a science concerned with the study of computational processes. A computational process is distinguished from, say, a chemical or electrical process, in that it is studied "in ways that ignore its physical nature" (Hailperin et al. 1999: 3). For example, the process by which a card player arranges cards in her hand, and the process by which a computer sorts names in a customer list, though they share nothing in common physically, may nevertheless embody the same computational process. They may, for example, both proceed by scanning the items to be arranged one by one, determining the proper place of each scanned item relative to the items already scanned, and inserting it into that place, perhaps necessitating the moving of previously scanned items to make room. This process (known as an insertion sort in computer science terms) can be precisely described in a formal language without talking about playing cards or semiconducting elements. When so described, one has a computational model of the process in the form of a computer program. This model can be tested, in a way analogous to how a hypothesis is tested in the natural sciences, by executing the program and observing its behavior. It can also be reasoned about abstractly, so that questions can be answered about it, such as whether there are other

processes which will have the same effect but achieve it more efficiently. Building computational models and answering these kinds of questions form a large part of what computer scientists do.

The explosive growth in the number of computer applications in the last several decades has shown that there is no limit to how many real-world processes are amenable to modeling by computer. Not only have traditional activities, like record-keeping, investing, publishing, and banking, been simply converted to control by computational models, but whole new kinds of activity have been created that would not be possible without such models. These are the by-now-familiar "virtual" activities described in the language of cyberspace: e-mail, chatrooms, web surfing, online shopping, internet gaming, and so on.

The role of philosophy in that subfield of computer science known as artificial intelligence (AI) has long been recognized, given the roles of knowledge and reasoning in AI. And the reverse, the role of AI in philosophy, has even been highlighted by some, albeit controversially. (See Chapter 9, The Philosophy of AI and its Critique, and Chapter 10, Computationalism, Connectionism, and the Philosophy of Mind.) But apart from considerations arising from the modeling of knowledge and reason, computer science is ripe for the good old-fashioned analysis that philosophy can provide for any science. Thus, a perfectly reasonable role of philosophy is to attempt to place computer science within the broad spectrum of inquiry that constitutes science. The concern here is to deal with the inevitable identity crises that crop up in the self-image of any adolescent, which computer science certainly is. Philosophy should address questions like: What is the relation between mathematics and computer science? Is there a sense in which computer science is experimental science? Is a computer programmer merely a data wizard, or can she also engage in information modeling? What is the nature of abstraction in computer science? What are the ontological implications of computer science concepts? From the point of view of computer science methodology, the most probing of these questions concerns the relation between mathematics and computer science and the nature of abstraction in computer science. The remainder of this chapter turns its attention to these issues.

Computer Science and Mathematics

Philosophical contributions to the foundations of scientific disciplines often center around "pivotal questions" regarding reductionist attempts. In the philosophy of biology, for example, the question is whether the science of the organic can be reduced to the science of the inorganic (the reduction of biological to physical laws). In mathematics, logicism claims that all of mathematics can be reduced to logic. For science in general, logical positivism advocated the reduction of theoretical vocabulary to observational vocabulary. An early "pivotal question" in the philosophy of computer science is whether computer science can be reduced to a branch of mathematics. How a computer scientist answers this question can greatly influence his or her methodology.

The range of perspectives from which the reductionist issue can be addressed is wide. Consider the following view, expressed by C. A. R. Hoare: "Computer programs are mathematical expressions. They describe, with unprecedented precision and in the most minute detail, the behavior, intended or unintended, of the computer on which they are executed" (Hoare 1986: 115). And this alternative, offered by C. Floyd: "Programs are tools or working environments for people. [They] are designed in processes of learning and communication so as to fit human needs" (Floyd 1987: 196). The view expressed by Hoare is unequivocal: computer programs are mathematical expressions. The quote by Floyd is less precise, but expresses a view on the function of programs for humans in decidedly non-mathematical terms. While these views do not necessarily contradict one another, they can most definitely signal contrasting interpretations as to how computer programs ought to be designed, built, and used.

While these quotations express views on the function and status of computer programs, the

differences of opinion extend to the broader notion of computing as a science, in which the task of actually creating a program is but one aspect. Of related concern, for example, are the activities of program specification and verification. A program specification is a detailed description of a program's input and output, ignoring the details of how the program actually accomplishes its task. Program verification is the process of determining whether a program actually conforms to its specification. These activities are just as germane to the software process as writing the program itself, and there is no agreement on whether or not program specifications should be mathematical entities and whether or not program verification can be a purely mathematical activity.

There is agreement, however, on the possibility of mathematically reasoning about programs as the *abstract* representations of algorithms, as opposed to programs as the causal manipulations of bits. For example, given a program P consisting of the statements $S_1; \ldots ; S_n$, it is possible to construct statements like "Let $S_1, S_2, \ldots ,$ and S_n be an abstract representation of program P. Then P has property R," where R describes some aspect of the execution of P in the abstract sense. For example, R might describe limits on the time it would take P to run, or the amount of memory P would require to execute. By giving precise interpretations to the S_i in a pertinent language and appropriately choosing R, it may be possible that the statement above is a theorem in a formal language. This is in fact the approach taken by modern researchers in computer science who are concerned with reasoning about algorithms and data structures.

While this is a justification of how reasoning *about* programs can be regarded as mathematical, it is yet a much broader claim to say that computer science is, or ought to aspire to be, a branch of mathematics. For there are still the issues of whether the specification, generation, or maintenance of programs (apart from reasoning about completed ones) is or ought to be like a mathematical activity. The issue which motivates and underlies much of the tension in philosophical discussion of computer science is formal verification, or mathematically reasoning about a program's outcome.

The Formal Verification Debate

The use of formal verification in computer science has generated debate since the appearance of a paper on verification and social processes by R. DeMillo, R. Lipton, and A. Perlis in 1979. But it was not until 1988 that these questions drew the attention of a "pure" philosopher, when J. Fetzer resurrected the program verification/social process debate of a decade earlier and subjected it to genuine philosophical analysis. Before this time, debate on the issues was evidenced mainly by differences in visionary accounts of how the young discipline of computer science ought to proceed, given not by philosophers but by computer science practitioners and teachers.

One of the early proponents of formal program verification was John McCarthy, who is also given credit for coining the term "artificial intelligence" in the 1950s. McCarthy was originally motivated by a theory of computation that would allow, among other advantages, the automatic translation from one linguistic paradigm to another. One can, in fact, look back now after nearly 30 years and confirm that automatic program translation, with the help of precise language specification, has been accomplished in the case of language compilers. These are programming tools that translate programs written in familiar human languages like Basic, C++, and Java, into the machine language of computers, which is composed only of zeroes and ones. However, as every language reference manual's warranty disclaimer demonstrates, no automatic compiler in all cases correctly translates programs into machine language. Thus, there is the distinction between (1) using mathematical methods during language translation to produce highly reliable machine language code, and (2) using mathematical methods to prove that a source program would behave, in an abstract sense, exactly as its specification implies. McCarthy, seeing no obstacle to (2), wrote:

> It should be possible almost to eliminate debugging. Debugging is the testing of a program on cases one hopes are typical, until it seems to work. This hope is frequently vain. Instead of debugging a program, one should prove that it meets its specifications, and this

> proof should be checked by a computer program. (McCarthy 1962: 22)

While McCarthy was one of the first to express this opinion, it came to be shared by others, who strove to describe what such a proof would be like. P. Naur recognized that one way to talk about a program both as a static, textual entity and as a dynamic, executing entity, was to conceive of the program as executing, but from time to time to conceptually "halt" it and make statements about the state of its abstract machine at the time of halting (Naur 1966). By making a number of these "snapshots" of a conceptually executing program, and by providing justifications for each on the basis of the previous one, a proof about the state of the abstract machine upon termination could be constructed.

Though this idea held much promise for believers in the mathematical paradigm (i.e., that computer science is a branch of formal mathematics), it came under attack in the abovementioned essay by DeMillo, Lipton, and Perlis. They argued that mechanically produced program verifications, which are long chains of dense logical formulas, are not what constitute mathematical proofs. In coming to be accepted, a mathematical proof undergoes social processes in its communication and peer scrutiny, processes that cannot be applied to unfathomable pages of logic. While DeMillo, Lipton, and Perlis did not subscribe to the mathematical paradigm, they also did not deny that programming is *like* mathematics. An analogy can be drawn between mathematics and programming, but "the same social processes that work in mathematical proofs doom verifications" (DeMillo et al. 1979: 275). Social processes, they argued, are critical:

> No matter how high the payoff, no one will ever be able to force himself to read the incredibly long, tedious verifications of real-life systems, and unless they can be read, understood, and refined, the verifications are worthless. (DeMillo et al. 1979: 276)

Although Fetzer was also a critic of the mathematical paradigm for computer science, it was for different reasons. He argued that the presence or absence of social processes is germane to neither the truth of mathematical theorems nor program verifications:

> Indeed, while social processes are crucial in determining what theorems the mathematical community takes to be true and what proofs it takes to be valid, they do not thereby make them true or valid. The absence of similar social processes in determining which programs are correct, accordingly, does not affect which programs are correct. (Fetzer 1988: 1049)

DeMillo, Lipton, and Perlis hit upon, for example, the boredom, tedium, and lack of glamor involved in reviewing proofs produced by mechanical verifiers. But for Fetzer, if this is all there is to their criticism of formal verification, it is not substantial. As Fetzer pointed out, social processes are characterized by transitory patterns of human behavior which, one could imagine, in different circumstances would reserve for program verification the same sort of excitement and collegial collaboration that marks the best mathematical research. Thus DeMillo, Lipton, and Perlis have identified a difference in *practice* between mathematical research and formal program verification, but not in *principle*.

Fetzer believes that formal program verification cannot fulfill the role that some of its advocates would assign to it within software engineering, but he attacks it from a nonsocial, more strictly philosophical perspective. This has to do with the relationship between mathematical models and the causal systems they are intended to describe. Close scrutiny of this relationship reveals, for Fetzer, the relative, rather than absolute, nature of the program correctness guarantee that formal verification can provide. It is only possible to prove formally something about a formal model, that is, a formal program model rendered in formal text. It is not possible to prove formally something about a causal model, that is, an actual, executing program represented in a physical, electronic substrate of bistable processor and memory elements. "[I]t should be apparent that the very idea of the mathematical paradigm for computer science trades on ambiguity" (Fetzer 1991: 209).

Strong claims of formal verificationists are victim to this ambiguity because they ignore several distinctions: between programs running

on abstract machines with *no* physical counterpart and programs running on abstract machines *with* a physical counterpart; between "programs-as-texts" and "programs-as-causes"; and between pure and applied mathematics. Recognizing these distinctions, for Fetzer, reveals that the claim that it is possible to reason in a purely *a priori* manner about the behavior of a program is true if the behavior is merely abstract; false, and dangerously misleading otherwise. This guarantees the indispensability of empirical methods in the software development process, for example, the use of program testing in an effort to eliminate program bugs.

Abstraction in Computer Science

Computer scientists are often thought to labor exclusively in a world of bits, logic circuits, and microprocessors. Indeed, the foundational concepts of computer science are described in the language of binary arithmetic and logic gates, but it is a fascinating aspect of the discipline that the *levels of abstraction* that one can lay upon this foundational layer are limitless, and make possible to model familiar objects and processes of everyday life entirely within a digital world. When digital models are sufficiently realistic, the environments they inhabit are called virtual worlds. So today, of course, there are virtual libraries, virtual shopping malls, virtual communities, and even virtual persons, like the digital version of actor Alan Alda created in an episode of PBS's *Scientific American Frontiers.*

Complex virtual worlds such as these are made possible by computer scientists' ability to distance themselves from the mundane and tedious level of bits and processors through tools of abstraction. To abstract is to describe something at a more general level than the level of detail seen from another point of view. For example, an architect may describe a house by specifying the height of the basement foundation, the location of load-bearing walls and partitions, the R-factor of the insulation, the size of the window and door rough openings, and so on. A realtor, however, may describe the same house as having a certain number of square feet, a certain number of bedrooms, whether the bathrooms are full or half, and so on. The realtor's description leaves out architectural detail but describes the same entity at a more general level, and so it is an abstraction of the architect's description. But abstraction is relative. For example, the architect's description is itself an abstraction when compared to a metallurgist's description of the nails, screws, and other fasteners making up the house, and the botanist's description of the various cellular properties of the wood it contains.

The computer scientist's world is a world of nothing but abstractions. It would not be possible to create the complex virtual worlds described above if the only things computer scientists could talk about were bits, bytes, and microcircuits. One can give an accurate idea of what computer scientists do by describing the abstraction tools they use. Now to characterize computer science as involved with abstractions seems to claim for it a place alongside mathematics as a purely formal endeavor. But the general trends in all programming are toward higher-quality software by abstracting away from the lower-level concepts in computer science and toward the objects and information that make up the real world. This is a kind of abstraction that is fundamentally different from that which takes place in mathematics. Understanding the difference is crucial in avoiding the persistent misconception by some that computer science is just a branch of pure mathematics. Both mathematics and computer science are marked by the introduction of abstract objects into the realm of discourse, but they differ fundamentally in the nature of these objects. The difference has to do with the abstraction of *form* versus the abstraction of *content.*

Traditionally, mathematics, as a formal science, has been contrasted with the factual sciences such as physics or biology. As natural sciences, the latter are not concerned with abstraction beyond that offered by mathematics as an analytical tool. The literature is full of strict bifurcations between the nature of formal and factual science in terms of the meanings of the statements involved in them. R. Carnap, for example, employs the analytic/synthetic distinction in claiming that the formal sciences contain only analytic statements. Since analytic statements are true only by virtue of the transformation rules of the language in

which they are made, Carnap is led to the view that "*[t]he formal sciences do not have any objects at all*; they are systems of auxiliary statements without objects and without content" (Carnap 1953: 128). Thus, according to Carnap the abstraction involved in mathematics is one totally away from content and toward the pure form of linguistic transformations.

Not all philosophers of mathematics agree with Carnap that mathematics has only linguistic utility for scientists, but there is agreement on the nature of mathematical abstraction being to remove the meanings of specific terms. M. Cohen and E. Nagel, for example, present a set of axioms for plane geometry; remove all references to points, lines, and planes; and replace them with symbols used merely as variables. They then proceed to demonstrate a number of theorems as consequences of these new axioms, showing that pure deduction in mathematics proceeds with terms that have no observational or sensory meaning. An axiom system may just *happen* to describe physical reality, but that is for experimentation in science to decide. Thus, again, a mathematical or deductive system is abstract by virtue of a complete stepping away from the content of scientific terms:

> Every [deductive] system is of necessity abstract: it is the structure of certain *selected* relations, and must consequently omit the structure of other relations. Thus the systems studied in physics do not include the systems explored in biology. Furthermore, as previously shown, a system is deductive not in virtue of the special meanings of its terms, but in virtue of the universal relations between them. The specific quality of the things which the terms denote do not, as such, play any part in the system. Thus the theory of heat takes no account of the unique sensory qualities which heat phenomena display. A deductive system is therefore doubly abstract: it abstracts from the specific qualities of a subject matter, and it selects some relations and neglects others. (Cohen & Nagel 1953: 138–9)

As a final example, consider C. Hempel's assessment of the nature of mathematics while arguing for the thesis of *logicism*, or the view that mathematics is a branch of logic:

> The propositions of mathematics have, therefore, the same unquestionable certainty which is typical of such propositions as "All bachelors are unmarried," but they also share the complete lack of empirical content which is associated with that certainty: The propositions of mathematics are devoid of all factual content; they convey no information whatever on any empirical subject matter. (Hempel 1953: 159)

In each of these accounts of mathematics, all concern for the content or subject-matter of specific terms is abandoned in favor of the *form* of the deductive system. So the abstraction involved results in essentially the *elimination* of content. In computer science, content is not totally abstracted away in this sense. Rather, abstraction in computer science consists in the *enlargement* of content. For computer scientists, this allows programs and machines to be reasoned about, analyzed, and ultimately efficiently implemented in physical systems. For computer users, this allows useful objects, such as documents, shopping malls, and chatrooms, to exist virtually in a purely electronic space.

Understanding abstraction in computer science requires understanding some of the history of software engineering and hardware development, for it tells a story of an increasing distance between programmers and the machine-oriented entities that provide the foundation of their work. This increasing distance corresponds to a concomitant increase in the reliance on abstract views of the entities with which the discipline is fundamentally concerned. These entities include machine instructions, machine-oriented processes, and machine-oriented data types. The remainder of this chapter will explain the role of abstraction with regard to these kinds of entities.

Language abstraction. At the grossest physical level, a computer process is a series of changes in the state of a machine, where each state is described by the presence or absence of electrical charges in memory and processor elements. But programmers need not be directly concerned with machine states so described, because they can make use of software development tools that allow them to think in other terms. For example, with the move from assembly to high-level language, computer scientists can abandon

```
for i ← 1 to n do
   for j ← 1 to m do
      read(A[i,j])
for j ← 1 to m do
   for k ← 1 to p do
      read(B[j,k])
for i ← 1 to n do
   for k ← 1 to p do begin
      C[i,k] ← 0
      for j ← 1 to m do
         C[i,k] ← C[i,k] + A[i,j] * B[j,k]
   end
```

Figure 24.1: Multiplying matrices

talk about particular machine-oriented entities like instructions, registers, and word integers in favor of more abstract statements and variables. High-level language programs allow machine processes to be described without reference to any particular machine. Thus, specific language content has not been eliminated, as in mathematical or deductive systems, but replaced by abstract descriptions with more expressive power.

Procedural Abstraction. Abstraction of language is but one example of what can be considered the attempt to enlarge the content of what is programmed about. Consider also the practice of *procedural abstraction* that arose with the introduction of high-level languages. Along with the ability to speak about abstract entities like statements and variables, high-level languages introduced the idea of *modularity*, according to which arbitrary blocks of statements gathered into *procedures* could assume the status of statements themselves. For example, consider the high-level language statements given in figure 24.1. It would take a studied eye to recognize that these statements describe a process of filling an $n \times m$ matrix A and an $m \times p$ matrix B with numbers and multiplying them, putting the result in an $n \times p$ matrix C such that $C_{i,k} = \Sigma_{j=1}^{m} A_{i,j}B_{j,k}$. But by abstracting out the three major operations in this process and giving them procedure names, the program can be written at a higher, and more readable, level as in figure 24.2. These three statements convey the same information about the overall process, but with less machine detail. No mention is made, say, of the order in which matrix elements are filled, or indeed of matrix subscripts at all. From the point

```
ReadMatrix(A,n,m)
ReadMatrix(B,m,p)
MultiplyMatrices(A,B,C,n,m,p)
```

Figure 24.2: Multiplying matrices with procedural abstraction

of view of the higher-level process, these details are irrelevant; all that is really necessary to invoke the process is the names of the input and output matrices and their dimensions, given as parameters to the lower-level procedures. Of course, the details of how the lower-level procedures perform their actions must be given in their definitions, but the point is that these definitions can be strictly separated from the processes that call upon them. What remains, then, is the total abstraction of a procedure's use from its definition. Whereas the language example had the abstraction of the content of computer instructions, here there is the abstraction of the content of whole computational procedures. And again, the abstraction step does not eliminate content in favor of form as in mathematics; it renders the content more expressive.

Data Abstraction. As a last example, consider the programmer's practice of *data abstraction.* Machine-oriented data types, such as integers, arrays, floating point numbers, and characters, are, of course, themselves abstractions placed on the physical states of memory elements interpreted as binary numbers. They are, however, intimately tied to particular machine architectures in that there are machine instructions specifically designed to operate on them. (For example, integer instructions on one machine may operate on 32 bits while similar operations on another machine may operate on 64 bits.) They are also built into the terminology of all useful high-level languages. But this terminology turns out to be extremely impoverished if the kinds of things in the world being programmed about include, as most current software applications do, objects like customers, recipes, flight plans, or chat rooms.

The practice of data abstraction is the specification of objects such as these and all operations that can be performed on them, without

reference to the details of their implementation in terms of other data types. Such objects, called *abstract data types* (ADTs), once they become implemented, assume their place among integers, arrays, and so on as legitimate objects in the computational world, with their representation details, which are necessarily more machine-oriented, being invisible to their users. The result is that programs that are about customers, recipes, flight plans, and so on are written in terms that are natural to these contexts, and not in the inflexible terms of the underlying machine. The programs are therefore easier to write, read, and modify. The specification and construction of abstract data types are primary topics in undergraduate computer science curricula, as evidenced by the many textbooks devoted to these topics. But this again is a type of abstraction that does not eliminate empirical content, as in mathematics, but rather enlarges the content of terms by bringing them to bear directly on things in a non-machine-oriented world.

Conclusion

Computer science will always be built on a scientific and engineering foundation requiring specialists with the most acutely analytic, creative, and technological minds. But the modeling and abstraction abilities that this foundation provides opens the field to virtually anyone willing to learn its languages. As computer science grows as a discipline, its methodology will be less dependent on the specialists maintaining the foundation and more dependent on those able to develop, implement, and use high-level languages for describing computational processes. What is meant by a "computational process" has diverged so much from the notion of a machine process that a currently popular language paradigm, namely object-oriented design and programming, de-emphasizes traditional algorithmic forms of program control in favor of the notions of classes, objects, and methods. (See, for example, the current widespread interest in Java.)

As programming languages evolve, it will be necessary for software developers to be conversant in the analytical tools of philosophers as they analyze their domains for logics, rules, classifications, hierarchies, and other convenient abstractions. Much of the computational modeling process will be characterized by activity more akin to logic and ontology than programming *per se*. (See the chapters in this volume by Smith, Antonelli, Gillies, and Bicchieri on ontology, logic, and probability.) Now more than ever, there is room for much philosophical research in the development of future computational modeling languages. One might even venture the prediction that philosophy will come to be an influential tool in the analysis and practice of computer science methodology.

References

Carnap, R. 1953. "Formal and factual science." In H. Feigl and M. Brodbeck, eds., *Readings in the Philosophy of Science*. New York: Appleton-Century-Crofts, pp. 123–8. [Carnap's paper is one of a group of seminal papers on the philosophy of mathematics collected in this anthology. They are central to understanding the distinction between mathematics and science, and the role mathematics plays in science. For the advanced philosophy student.]

Cohen, M. and Nagel, E. 1953. "The nature of a logical or mathematical system." In H. Feigl and M. Brodbeck, eds., *Readings in the Philosophy of Science*. New York: Appleton-Century-Crofts, pp. 129–47. [This paper is from the same group as Carnap 1953.]

DeMillo, R., Lipton, R., and Perlis, A. 1979. "Social processes and proofs of theorems and programs." *Communications of the ACM* 22: 271–80. [This paper ignited a debate within the computer science community regarding the role of formal methods in computer science methodology, specifically program verification. For the intermediate computer science student.]

Fetzer, J. 1988. "Program verification: the very idea." *Communications of the ACM* 31: 1048–63. [This paper rekindled the program verification debate from a philosopher's point of view and caused impassioned responses from the computer science community. For the intermediate philosophy student.]

——. 1991. "Philosophical aspects of program verification." *Minds and Machines* 1: 197–216. [This

paper summarized the wreckage of the program verification debate, again from the philosophical point of view, arguing against the mathematical paradigm. For the intermediate philosophy student.]

Floyd, C. 1987. "Outline of a paradigm change in software engineering." In G. Bjerknes et al., eds., *Computers and Democracy: A Scandinavian Challenge*. Hants., England: Gower, pp. 191–210. [This paper was one of the first to come out of the computer science community advocating a view of software as process rather than product. For the introductory student in any discipline.]

Hailperin, M., Kaiser, B., and Knight, K. 1999. *Concrete Abstractions: An Introduction to Computer Science*. Pacific Grove, CA: PWS Publishing. [An excellent introduction to computer science that emphasizes the role of abstractions in computer science methodology. For the first time computer science student.]

Hempel, C. 1953. "On the nature of mathematical truth." In H. Feigl and M. Brodbeck, eds., *Readings in the Philosophy of Science*. New York: Appleton-Century-Crofts, pp. 148–62. [This paper is from the same group as Carnap 1953.]

Hoare, C. 1986. "Mathematics of programming." *BYTE*, Aug.: 115–49. [A clear statement of the view that computer science is a species of pure mathematics. For the advanced computer science student.]

McCarthy, J. 1962. "Towards a mathematical science of computation." *Proceedings of the IFIP Congress* 62: 21–8. [Another statement of the mathematical paradigm for computer science, with an emphasis on the role of recursion in computer science methodology. For the advanced computer science student.]

Naur, P. 1966. "Proof of algorithms by general snapshots." *BIT* 6: 310–16. [An example of mathematical reasoning about programs in pursuit of program verification. For the advanced computer science student.]

Chapter 25

Philosophy of Information Technology

Carl Mitcham

Philosophy of information technology may be seen as a special case of the philosophy of technology. Philosophical reflection on technology aims in general to comprehend the nature and meaning of the making and using, especially of things made and used. Such reflection nevertheless exhibits a tension between two major traditions: one arising within engineering, another in the humanities (Mitcham 1994). For the former or expansionist view, technology is deeply and comprehensively human, and thus properly extended into all areas of life; according to the latter or limitationist perspective, technology is a restricted and properly circumscribed dimension of the human. This distinction and corresponding tensions may also be seen at play in the philosophy of information technology (IT), between those who would critically celebrate and extend IT and those who would cautiously subordinate and delimit it. Diverse metaphysical, epistemological, and ethical arguments are marshaled to defend one position over the other, as well as to build bridges between these two philosophical poles.

Philosophies of *x* commonly begin with attempts to define *x*. Philosophy of science, for instance, logically opens with the demarcation problem, by considering various proposals for distinguishing science from other forms of knowledge or human activity. The philosophy of information technology, like the philosophy of computer science, is properly initiated by the effort to define that on which it seeks to reflect. Once preliminary definitions are negotiated, philosophies of *x*, often against a historico-philosophical background, recapitulate in differentially weighted forms the main branches of philosophy *tout court* – metaphysics, epistemology, and ethics – with particular emphases reflecting both the unique philosophical challenges of *x* and the context of presentation. In the present case, for example, although ethical and political issues play a prominent role in the philosophy of information technology, they are treated lightly here because of the more extensive coverage provided by Chapters 5 and 6 (Computer Ethics and Computer-mediated Communication and Human–Computer Interaction). Here the stress is on theoretical issues concerning especially metaphysical assessments of information technology.

What Is Information Technology?

Information technology – or such closely related terms as "information systems" and "media technology" – is commonly described as that

technology constituted by the merging of data-processing and telecommunications (with diverse input devices, processing programs, communications systems, storage formats, and output displays). It arose from earlier forms of electronic communications technology (telegraph, telephone, phonograph, radio, motion pictures, television) by way of computers and cybernetics (see Chapter 14), an earlier term that still casts its shadow over IT, as in such coinages as "cyberspace" and other cognates. It may nevertheless be useful to begin by attempting to rethink what is perhaps too facile in such a description.

The terms "information" and "technology" are both subject to narrow and broad, not to say engineering and humanities, definitions. Developed by Claude Shannon (Shannon & Weaver 1949), the technical concept of information is defined as the probability of a signal being transmitted from device A to device B, which can be mathematically quantified (see Chapters 4 and 13, treating Information and The Physics of Information, respectively). The theory of information opened up by this distinct conceptual analysis has become the basis for constructing and analyzing digital computational devices and a whole range of information (also called communication) technologies, from telephones to televisions and the internet.

In contrast to information (and information technologies) in the technical sense is the concept of information in a broader or semantic sense. Semantic information is not a two-term relation – that is, a signal being transmitted from A to B – but a three-term relation: a signal transmitted from one device to another, which is then understood as saying something to a person C. Although information technologies in the technical sense readily become information technologies in the semantic sense, there is no precise relation between technical and semantic information. Independent of its probability as a signal, some particular transmission may possess any number of different semantic meanings. A signal in the form of two short clicks or light flashes (Morse Code for the letter "i"), could be a self-referential pronoun, part of the word "in," a notation in Latin numerals of the number one – or any number of other possibilities. Absent the context, a signal is not a message. Kenneth Sayre (1976) and Fred I. Dretske (1981) are nevertheless two important attempts to develop semantic theories of information grounded in the technical concept of information (see Chapter 17, Knowledge).

"Technology" too is a term with narrow and broad definitions. In the narrow or engineering sense, technology is constituted by the systematic study and practice of the making and using of artifacts (cf. the curricula of technological universities), and to some extent by the physical artifacts themselves (from hammers to cars and computers). Indeed, a distinction is often drawn between premodern *techne* or technics and modern technology. For thousands of years human making and using proceeded by intuitive, trial-and-error methods, remained mostly small-scale and dwarfed by natural phenomena. With the rise of modern methods for making and using, these activities became systematically pursued (often on the basis of scientific theories) and their products began to rival natural phenomena in scale and scope. In a broader humanities parlance, technology covers both intuitive, small-scale and scientific, large-scale making and using in all its modes – as knowledge, as artifact, as activity, and even as volition.

Given these narrow and broad definitions for each element in the compound term, one may postulate a two by two matrix and imagine four different information technology exemplars (figure 25.1). In what follows, a significant sample from among these possible information technologies will be analyzed in order to illustrate diverse facets of a potentially comprehensive philosophy of information technology.

Information Technology in Historico-philosophical Perspective

Philosophy is not coeval with human thought, but emerges from and against prephilosophical reflection that it nevertheless continues or mirrors. Prior to the rise of philosophy, mythological and poetic narratives often expressed the ambivalence of the human experience of tool making

	Premodern technology	*Scientific technology*
Technical sense of information	Alphabetic writing	Electronic and source code signal transmission
Semantic sense of information	Books and related texts	Works of high representational electronic communications media (movies, TV programs, hypertexts, etc.)

Figure 25.1: Information technology exemplars

and using. Stories of the conflict between Cain (builder of cities) and Abel (pastoral shepherd), of Prometheus (who stole fire for humans from the gods), of Hephaestus (the deformed god of the forge), and of Icarus (the inventor who went too far) all attest to the problematic character of human engagement with what has come to be called technology. The story of the Tower of Babel (Genesis 11) even suggests the destructive linguistic repercussions of an excessive pursuit of technological prowess.

By contrast, when the prophet Ezekiel learns in the desert to infuse dry bones (alphabetic consonants) with the breath of the spirit (unwritten vowels), it is as if God were speaking directly through him (Ezekiel 37). Indeed, God himself creates through speech or *logos* (Genesis 1), and writes the law both in stone and in the hearts of a people. Thus, information technologies in their earliest forms – speech and writing – manifest at least two fundamental experiences of the human condition: sin or hubris and transcendence, the demonic and the divine.

Greek philosophical reflection on *techne* likewise noted the two-fold tendency of human skill in the making and using of artifacts to be pursued in isolation from the good and to participate in the divine. This is as true of information *technai*, such as oratory and writing, as it is of the mechanical and military arts. In Plato's *Gorgias*, for instance, Socrates challenges the sophist to reintegrate the techniques of rhetoric with the pursuit of truth, to eschew the tricks of gaining power divorced from knowledge of the good. In the *Phaedrus*, Socrates tells the story of how King Thamus rejected the Egyptian god Theuth's invention of writing on the grounds that it would replace real with merely virtual memory (*Phaedrus* 274d ff.). Socrates himself comments on the silence of written words, and Plato famously remarked on the limitations of writing even in his own works (*Letter* VII, 341b–e and 344c–d). The *Politicus* (300c ff.), however, concludes with a modest defense of written laws, and the *Ion* presents the poet as one inspired by the gods.

Aristotle, in an analysis that echoes Plato's assessment in *The Republic* of artifice and poetry as thrice-removed from being itself, notes the inability of *techne* to effect a substantial unity of form and matter. "If a bed were to sprout," says Aristotle, "not a bed would come up but a tree" (*Physics* II, i, 193a12–16). In a parallel analysis of the relation between experiences, spoken words, and writing at the beginning of *On Interpretation*, Aristotle places the written word at two removes from experience and three removes from the things experienced, thus implying a dilution of contact with reality as one moves from the information technology of speech to that of writing. Spoken words refer to experience; written words to spoken words.

In contrast to Aristotle's characterization of words in strictly human terms, Christianity reaffirms the divine character of the transcendent word incarnate (John 1) and of the transmission of the gospel through that preaching which represents the word (Romans 10:17). Indeed, according to Augustine, Christian preaching unites truth and language with an efficacy that the Platonists could not imagine (*De vera religione* i, 1 ff.). This is an argument that has been revived in Catherine Pickstock's theological interpretation of that information technology known as liturgy (Pickstock 1998). At the same time, the meaning of the words of revelation in Scripture is not

always obvious, thus requiring the development of principles of interpretation (see Augustine's *De doctrina christiana*). Faith in the Scriptures as the word of God solves, as it were, the technical question concerning the extent to which the signal has been accurately transmitted from A to B (God to humans), but not the semantic question of what this signal means (to whom it speaks and about what). The meaning of revelation requires a science of interpretation or hermeneutics if its information (from the Latin *informare*, to give form) is truly to convert those who receive it.

This dedication to the development of techniques of interpretation led to a unique medieval flowering of logical, rhetorical, and hermeneutic prowess. Reflecting the effulgence of poetic exegesis of sacred texts, Thomas Aquinas defends the metaphorical "hiding of truth in figures" as fitting to the word of God, and argues the power of Scripture to signify by way of multiple references: historical or literal, allegorical, tropological or moral, and anagogical or eschatological (*Summa theologiae* I, q.1, art.9–10). What is equally remarkable is that – no doubt stimulated by the literal and spiritual interpretations of revelation as granting the world a certain autonomy and calling upon human beings to exercise positive mastery over it – the flowering of semantic studies was paralleled by an equally unprecedented blossoming of physical technologies. Examples include the waterwheel and windmill, the moldboard plow, the horse collar, the lateen sail, and the mechanical clock.

The modern world opens, paradoxically, by pitting the second form of technological progress (physical inventions) against the first (poetic creativity). Metaphorical words are to be rejected in the pursuit of real things and ever more powerful technologies (see especially the arguments of Francis Bacon and René Descartes). The historical result was to turn exegesis into criticism and semantic analysis into a drive for conceptual clarity, in a reform of the techniques of communication that became most manifest in the new rhetoric of modern natural science – as well as in the invention of a whole new information technology known as moveable type.

The invention of the printing press and the consequent democratization of reading can be associated with a manifold of social transformations: religious, political, economic, and cultural. The philosophical influences of such changes have been legion. To cite but one example, as the world was increasingly filled with texts, and texts themselves were severed from stable lifeworlds of interpreters, philosophy became increasingly linguistic philosophy, in two forms. In continental Europe, hermeneutics was redefined by Friedrich Schleiermacher as the interpretation of all (not just sacred) texts, by Wilhelm Dilthey as the foundation of the *Geistwissenschaften* or humanistic sciences, and by Martin Heidegger as the essence of *Dasein* or human being. In this same milieu, Ferdinand de Saussure invented the science of linguistics, focusing neither on efficient signal transmission nor on multiple levels of external reference but on language as a system of words that mutually define one another through their internal relations. In the Anglo-American world, especially under the influence of Ludwig Wittgenstein, philosophy became linguistic philosophy, which takes the meaning of words to be constituted by their uses, thus calling attention to multiple contexts of use, what Wittgenstein called ways of life. Indeed, in some forms the resultant philosophy of language turns into a kind of behaviorism or is able to make common cause with pragmatism.

In another instance, theories were posited about the relation between changes in information technologies and cultural orders. The contrast between orality and literacy has been elaborated by a series of scholars – from Albert Lord and Milman Parry to Marshall McLuhan, Walter Ong, and Ivan Illich – who have posited complementary theories about relations between information technology transformations and cultures. With McLuhan, for example, there is a turn not just from technical signal to semantic message, but an attempt to look at the whole new electronic signal transmitting and receiving technology (never mind any specific semantic content) as itself a message. In his own condensed formulation: the medium (or particular form of information technology) is the message (McLuhan 1964).

Stimulated especially by McLuhan, reviews of the historical influences between philosophy and IT begin to mesh into a philosophy of history

that privileges IT experience the way G. W. F. Hegel privileged politics and ideas. Here Paul Levinson's "natural history of information technology" (1997) is a worthy illustration.

Information Technology and Metaphysics

Although the historico-philosophical background points to an emergence, in conjunction with information technology, of new cultural constellations in human affairs, pointing alone is insufficient to constitute philosophy. Popular attempts to think the new IT lifeworld have emphasized economics and politics, in which issues are decided about e-banking and e-commerce on the basis of market forces and political power. The ethics of information technology, as an initiation into philosophical reflection – that is, into thought in which issues are assessed on the basis of argument and insight rather than money and votes – has highlighted issues of privacy, equity, and accountability. Yet given that the fundamental question for ethics concerns how to act in accord with what really is, there are reasons to inquire into the kind of reality disclosed by IT – that is, to raise metaphysical (beyond the physical or empirical) and ontological (from *ontos*, the Greek word for "being") questions.

What are the fundamental structures of the IT phenomenon? What is real and what is appearance with regard to IT? Richard Coyne (1995), for instance, argues that it is illusory to view IT as simply a novel instrument available for the effective realization of traditional projects for conserving and manipulating data. Albert Borgmann (1999) insightfully distinguishes between information about reality (science), information for reality (engineering design), and information as reality (the high-definition representations and creations emerging from IT) – and further the increasing prominence, glamor, and malleability of information as reality is having the effect of diminishing human engagement with more fundamental realities. With regard to the kinds of metaphysical issues raised by Coyne, Borgmann, and others, it is useful to distinguish again expansionist and limitationist approaches to the nature and meaning for information technology.

The expansionist approach has its roots in technical thinking about IT, first in terms of physical entities. At least since Norbert Wiener (1948) effectively posited that, along with matter and energy, information is a fundamental constituent of reality, questions have been raised about the metaphysical status of information. Building on Wiener's own analysis, distinctions may be drawn between three fundamentally different kinds of technology: those which transform matter (hammers and assembly lines), those which produce and transform energy (power plants and motors), and those which transform information (communication systems and computers).

A related phenomenology of human engagement would observe how the being of IT differs from tools and machines. Unlike tools (which do not function without human energy input and guidance) or machines (which derive energy from nonhuman sources but still require human guidance), information technologies are in distinctive ways independent of the human with regard to energy and immediate guidance; they are self-regulating (cybernetic). In this sense, steam engines with mechanical governors on them or thermostatically controlled heating systems are examples of information machines. Insofar as the operation of more electronically advanced IT is subject to human guidance, guidance ceases to be direct or mechanical and is mediated by humanly constructed programs (electronically coded plans). What is the ontological status of programs? What are their relations to intentions? Indeed, in IT, operation and use appear to have become distinguishable. IT is a new species of artifact, a hybrid that is part machine running on its own and part utility structure like a road waiting to be driven on – hence the term "new media" (as both means and environment). The static availability of such structures is contingent on their semi-autonomous dynamic functioning.

Second, in terms of the cognitive capabilities of IT, transempirical questions arise about the extent to which computers (as pervasive elements in IT) imitate human cognitive processes. Do computers think? What kind of intelligence is artificial intelligence (AI)? Are the different kinds

of AI – algorithmic, heuristic, connectionist, embodied, etc. – different forms of intelligence? Such ontological questions now blend into others, concerning the extent to which high-tech artifacts are different from living organisms. Biotechnology has breached Aristotle's distinction between natural tree and artificial bed, growth and construction, the born and the made. Soon computer programs may also be able not just to mimic patterns of growth on the screen (artificial life, see Chapter 15), but autonomous, artificial agents that are able to reproduce themselves. At the nano-scale, robotic design will hardly be distinct from genetic engineering. Will any differences in being remain?

From the technical perspective, information is ubiquitous in both the organic and the artificial worlds. The wall between the two is vanishing, although, insofar as the technical concept of information becomes a category of explanation in biology, it has also been argued to have distinct ideological roots (see Kay 2000 on this point). The cyborg (cybernetic organism) is a living machine, not a goddess (Haraway 1991). Within such a reality, the ethical imperative becomes experimenting with ourselves, what Coyne (1995) calls a pragmatic interaction with advancing IT. This is an attitude widely present among leading IT designers such as Mark Weiser at the famous Xerox Palo Alto Research Center (PARC), the ethos of which is commonly celebrated in *Wired* magazine. It has also have been given philosophical articulation by media philosopher Wolfgang Schirmacher. For Schirmacher (1994), IT is a kind of artificial nature, a post-technology in which we are free (and obligated, if we would act in harmony with the new way of being in the world) to live without predeterminations, playfully and aesthetically.

The limitationist approach originates in a different, more skeptical stance. Issues are no doubt oversimplified by characterizing one approach as pro-IT and another as con-IT – although such a contrast captures some measure of real difference (but see Gordon Graham, 1999, for a down-to-earth philosophical utilization of this contrast using the terms technophiles and neoluddites). Perhaps a better contrast would be that of Hegel versus Socrates: the comprehensive critical affirmation as opposed to the argumentative gadfly. From the Hegelian perspective there is something both adolescent and irresponsible about an ongoing Socratic negativity that refuses to take responsibility for world creation. Indeed, Socratic negativity easily becomes a philosophically clichéd substitute for true thinking. From the Socratic perspective, however, the expansionist approach comes on the scene as a court philosophy, especially insofar as it flatters the king and counsels expanding an already popular and widely affirmed domain of influence. In a state already dominated by information technology, the Socratic tradition thus finds expression in repeatedly questioning the nature and meaning of IT – a questioning that must ultimately go metaphysical.

At a first level, however, the questioning of IT will be, as already suggested, ethical. For instance, does IT not threaten privacy? Even more profoundly, does the IT mediation of human action in complex software programs, which are created by multiple technicians and are not even in principle able to be fully tested (Zimmerli 1986), not challenge the very notion of moral accountability? At a second level are political questions: Is the internet structured so as to promote social justice through equity of access? Is it compatible with democracy? Furthermore, IT exists on the back of a substantial industrial base, whose environmental sustainability is at least debatable. Insofar as IT depends on an unsustainable base, might not its own justice and goodness be compromised? At still a third level are psychological questions, blending into epistemological ones. Does the exponential growth of information availability not challenge the human ability to make sense of it? Information overload or information anxiety (see Wurman 2001) is one of the most widely cited paradoxes of IT life. Finally, at a fourth level are psychological-anthropological questions about the social implications of the new "mode of information" (Poster 1990), what it means to live a "virtual life" (Brook & Boal 1995) and "life on the screen" (Turkle 1995).

The third and fourth dimensions of limitationist, Socratic questioning – that is, the epistemological and anthropological levels – hint at the metaphysical. Information technology may hide reality from us in a much more fundamental way

than simply by means of information overload. It may deform our being at deeper levels than the psychological. To develop this possibility it is useful to refer at some length to Martin Heidegger, the most influential exponent of this position.

According to Heidegger's highly influential argument in "The Question Concerning Technology" (1977 [1954]), technology is constituted not so much by machines or even instrumental means in general as by its disclosure of reality, its unhiding, its truth. Premodern technology in the form of *poiesis* functioned as a bringing or leading forth that worked with nature, and as such revealed Being as alive with its own bringing forth, the way a seed blossoms into a flower or an acorn grows into an oak tree. Modern technology, by contrast, is not so much a bringing forth as a challenging-forth that reveals the world as *Bestand* or manipulatable resource.

In reading Heidegger it is crucial to recognize that he felt it necessary to couch his insights in a special vocabulary ("bringing forth," "challenging forth," "*Bestand*"), because of the way ordinary concepts are sedimented with assumptions that themselves help conceal the dimensions of reality to which he invites attention. In Heidegger's own words:

> The revealing that rules throughout modern technology has the character of a setting-upon, in the sense of a challenging-forth. That challenging happens in that the energy concealed in nature is unlocked, what is unlocked is transformed, what is transformed is stored up, what is stored up is, in turn, distributed, and what is distributed is switched about ever anew. Unlocking, transforming, storing, distributing, and switching about are ways of revealing. (Heidegger 1977 [1954]: 297–8)

To this distinctive way of revealing Heidegger also gives a special name: *Gestell* or enframing.

Although Heidegger seems to be thinking here of electric power generation, the same description would in many ways be applicable to information technology. There is a challenging that happens when digitally concealed information is unlocked (from, say, a computer disk), transformed (by some software program), stored up (on a hard drive), distributed (by internet), and switched about (forwarded, reprocessed, data mined, etc.). Indeed, in another text Heidegger makes the reference to IT explicit, although under the name of cybernetics. "Cybernetics," he writes, "transforms language into an exchange of news. The arts become regulated-regulating instruments of information" (Heidegger 1977 [1966]: 376). Modern information technology thus does to language what modern non-information technology does to the material world: turns it into *Bestand*, that is, a resource for human manipulation.

What is wrong with this? The basic answer is that modern technology, including modern information technology, conceals as well as it reveals. Insofar as we persist in emphasizing the revealing and ignore the concealing, concealing will actually dominate. We will not be fully aware of what is going on. To develop this point requires a brief elaboration of Heidegger's theory of hermeneutics. In his version of hermeneutics, which argues interpretation (more than rationality) as the defining characteristic of the human, Heidegger makes two basic claims.

The first is that no revealing (the acquisition of information in the semantic sense) is ever simple; it always involves the process of interpretation. Interpretation itself proceeds in texts, in perception, in thinking, and in life by means of a dialectic between part and whole, what is called the hermeneutic circle. The part is only revealed in terms of the whole, and the whole in terms of the parts. As a result, Heideggerian hermeneutics postulates a pregivenness in all revealing or, as he also likes to say, unconcealing. Our minds and our lives open not as with a *tabula rasa*, but with an immanent reality (both part and whole) waiting to be brought forth into the light of appearances. Understanding proceeds by means of a process of moving from part to whole and *vice versa*, repeatedly to make the implicit explicit, to reveal the concealed, analogous especially to the ways that premodern technology also worked to till the fields and to fashion handcrafted artifacts. The upshot is that not only is all information subject to interpretation, but that all information technology is part of a larger lifeworld and cannot be understood apart from such an implicit whole. To think otherwise is a metaphysical mistake.

The second claim is that any unconcealing is at one and the very same time a concealing. This second claim has even more profound implications for information technology, which through its expanding realms makes information more and more omnipresent. Information technology appears to reveal with a vengeance. According to Heidegger, however, this is ultimately an illusion – and dangerous to what it means to be human. The problem is not just one of sensory or information overload, but of information as a concealing of Being itself, the fundamental nature of reality, of the distinctly human relation to such reality.

For Heidegger the rise of modern technology, and its culmination in cybernetics or information technology, is the culmination of a historico-philosophical trajectory of thinking that began with the Greeks. With Plato and Aristotle, Being was first revealed, however tentatively and minimally, as a presence that could be re-presented in thought or rationalized. Over the course of its 2,500-year history, philosophy has successively spun off the various scientific disciplines as specialized ways to re-present the world: in mathematics, in logic, in astronomy, in physics, in chemistry, in biology, in cosmology, and now in the interdisciplinary fields of molecular biology, cognitive science, and more. This continuing development is the end of philosophy in two senses: its perfection and its termination. The very success of scientific revealing grew out of a specialization of thinking as philosophy that entailed leaving behind or concealing thinking in a more fundamental sense, something that Heidegger refers to as *Lichtung*, translated variously as "lighting" or "opening." "Perhaps there is a thinking," Heidegger writes, "which is more sober-minded than the incessant frenzy of rationalization and the intoxicating quality of cybernetics" (Heidegger 1977 [1966]: 391).

In another text, Heidegger describes this "new task of thinking" at "the end of philosophy" by means of a comparison between what he calls calculative and meditative thinking. "Calculative thinking never stops, never collects itself. Calculative thinking is not meditative thinking, not thinking which contemplates the meaning which reigns in everything that is" (Heidegger 1966 [1955]L 46). Meditative thinking, premodern and even preclassical Greek philosophical thinking, which was once in touch with the root of human existence, and out of which by means of a narrowing and intensified calculative thinking has emerged, has been replaced by calculative thinking in the form of "all that with which modern techniques of communication stimulate, assail, and drive human beings" (Heidegger 1966 [1955]: 48). Technology, especially information technology, conceals this meditative thinking, which Heidegger terms *Gelassenheit*, releasement or detachment. "Releasement toward things and openness to the mystery . . . promise us a new ground and foundation upon which we can stand and endure the world of technology without being imperiled by it" (Heidegger 1966 [1955]: 55). The fundamental threat in information technology is thus a threat to the human being's "essential nature" and the "issue of keeping meditative thinking alive" (ibid.: 56).

Current Research and Open Issues

What is most remarkable is the fact that Heidegger's radical critique of technology in general and information technology in particular has been subject to significant practical appropriations by IT users and designers, thus building bridges between the engineering and humanities, the expansionist and limitationist, traditions in the philosophy of information technology. Raphael Capurro (1986), for instance, brings Heidegger to bear on the field of library and information science. Hubert Dreyfus (2001) examines the Internet from a philosophical perspective indebted to Heidegger. With slightly more expansion, one may also reference two other leading examples: Terry Winograd and Fernado Flores, and Richard Coyne. At the same time serious challenges have been raised by Mark Poster to the adequacy of a Heideggerian approach to IT.

In the mid-1980s, computer scientists Winograd and Flores argued at length that Heideggerian analyses could disclose the reasons behind the failures of information technologies to function as well in the office as computer

scientists predicted. In Winograd and Flores (1987) they argue that Heideggerian insights can thus be a stimulus for redesigning computer systems.

A decade later architectural theorist Coyne (1995) goes even further, arguing that not just Heidegger but the post-Heideggerian thought of Jacques Derrida provides a philosophical account of what is going on among leading-edge information technology designers. Building on Heidegger's notion that all revealing involves a simultaneous concealing, Derrida proposes to deconstruct specific concepts, methods, and disciplinary formations precisely to bring to light their hidden aspects, that on which they depend without knowing or acknowledging it. For Coyne this opens the way for and justifies the turn from a commitment to rational method in information technology design to the renewed reliance on metaphor.

Heidegger and Derrida thus revalidate the creative significance of metaphor – of thinking of a computer operating system as "windows," of a screen "desktop" with "icons," even of the mind as a computer. It is precisely a play with such "irrational" connections that facilitates advances in information technology design. With Aquinas, Coyne seeks to defend the metaphorical "hiding of truth in figures" as functional not just in theology but also in technology. Whether either Aquinas or Heidegger would counsel such appropriation of their philosophies of information technology is, of course, seriously in doubt.

As if to reinforce such doubt about such creative appropriations, Poster argues at length that Heidegger "captures the revealing of modern technology only, not postmodern technology." Indeed, "some information technologies, in their complex assemblages, partake not only of [*Gestell*] but also of forms of revealing that do not conceal but solicit participants to a relation to Being as freedom" (Poster 2001: 32–3). For Poster a more adequate approach to the philosophical understanding of IT is through Felix Guattari's image of the rhizome and a phenomenology of the enunciative properties of specific technologies. A potentially comprehensive philosophy of IT thus remains, not unlike all philosophy, suspended in and energized by its fundamental alternatives.

References

Borgmann, A. 1999. *Holding On To Reality: The Nature of Information at the Turn of the Millennium*. Chicago: University of Chicago Press. [Distinguishes natural information (about reality), cultural information (for constructing reality), and technological information (information becoming a reality in its own right). Seeks to establish guidelines for assessing and limiting information as reality.]

Brook, J. and Boal, I. A. 1995. *Resisting the Virtual Life: The Culture and Politics of Information*. San Francisco: City Lights. [Twenty-one critical essays on IT inequities, impacts on the body, the degrading of the workplace, and cultural deformations.]

Capurro, R. 1986. *Hermeneutik der Fachinformation*. Freiburg: Alber. [For a short English paper that reviews the thesis of this book, see R. Capurro, "Hermeneutics and the phenomenon of information," *Research in Philosophy and Technology* 19: 79–85.]

Coyne, R. 1995. *Designing Information Technology in the Postmodern Age: From Method to Metaphor*. Cambridge, MA: MIT Press. [Advances in information technology are examined from the diverse philosophical perspectives of analytic philosophy, pragmatism, phenomenology, critical theory, and hermeneutics, in order to reveal their different implications for the design and development of new electronic communications media.]

Dretske, F. 1983. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press. [The most well-developed theory of perception and empirical knowledge based on mathematical information theory.]

Dreyfus, H. L. 2001. *On the Internet*. New York: Routledge. [A phenomenologically influenced but interdisciplinary critique.]

Graham, G. 1999. *The Internet: A Philosophical Inquiry*. New York: Routledge. [Questions and criticizes both neoluddite and technophile claims about the dangers and implications of the internet.]

Haraway, D. 1991. *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge. [See especially the "Manifesto for Cyborgs" included in this book.]

Heidegger, M. 1966 [1955]. *Discourse on Thinking*, tr. J. M. Anderson and E. H. Freund. New York: Harper and Row. [This includes translation of Heidegger's essay, "Gelassenheit."]

——. 1977 [1954]. "The question concerning technology," tr. W. Lovitt. In M. Heidegger, *Basic Writings*. New York: Harper and Row, pp. 287–317. [Heidegger's most important critique of technology.]

——. 1977 [1966]. "The end of philosophy and the task of thinking," tr. J. Stambaugh. In M. Heidegger, *Basic Writings*. New York: Harper and Row, pp. 373–92. [A brief statement of Heidegger's philosophy of the history of philosophy.]

Kay, L. E. 2000. *Who Wrote the Book of Life: A History of the Genetic Code*. Stanford: Stanford University Press. [A critical assessment of information as a metaphor in biology.]

Levinson, P. 1997. *The Soft Edge: A Natural History and Future of the Information Revolution*. New York: Routledge. [A new information medium (such as computers) does not so much replace an old medium (such as the telephone) as complement it.]

McLuhan, M. 1964. *Understanding Media: The Extensions of Man*. New York: McGraw-Hill. [It is not the information content of a medium (such as speech or television) that is most influential on a culture, but the character or structure of the medium itself. Electronic media are structurally distinct from, say, books. "The medium is the message."]

Mitcham, C. 1994. *Thinking through Technology: The Path between Engineering and Philosophy*. Chicago: University of Chicago Press. [A general introduction to the philosophy of technology that distinguishes two major traditions: engineering and humanities philosophy of technology. The former argues for the expansion, the latter for the delimitation of technology as object, knowledge, activity, and volition.]

Pickstock, C. 1998. *After Writing: On the Liturgical Consummation of Philosophy*. Oxford, UK: Blackwell. [A critique of Derrida and defense of information as subordinate to the context created by linguistic and bodily performance in a historical tradition.]

Poster, Mark. 1990. *The Mode of Information: Poststructuralism and Social Context*. Chicago: University of Chicago Press. [Argues that four new modes of information – TV ads, data bases, electronic writing, and computer science – create a world in which humans are socially constituted differently than in pre-electronic IT history.]

Poster, Mark. 2001. *What's the Matter with the Internet*. Minneapolis, MI: University of Minnesota Press. [A critique of applying Heidegger's analysis of technology to information technology, with special reference to postmodern thinkers such as Felix Guattari.]

Sayre, K. M. 1976. *Cybernetics and the Philosophy of Mind*. London: Routledge & Kegan Paul. [Argues a naturalist theory of mind based on mathematical information theory.]

Schirmacher, W. 1994. "Media and postmodern technology." In G. Bender and T. Druckrey, eds., *Culture on the Brink: Ideologies of Technologies*. Seattle: Bay Press. [The other contributions to this book are useful as well.]

Shannon, C. and Weaver, W. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press. [Contains two classic papers: Shannon's, from the *Bell System Technical Journal* (1948); and Weaver's, from *Scientific American* (1949).]

Turkle, Sherry. 1995. *Life on the Screen: Identity in the Age of the Internet*. New York: Simon & Schuster. [A psychologist's analysis of emerging forms of self-definition unique to the internet experience.]

Wiener, N. 1948. *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press. [The classic statement of the engineering theory of cybernetics. In other works Wiener also examined the social and ethical implications of his theories.]

Winograd, T. and Flores, F. 1987. *Understanding Computers and Cognition: A New Foundation for Design*. Reading, MA: Addison-Wesley. [A Heideggerian analysis by two computer scientists.]

Wurman, R. S. 2001. *Information Anxiety 2*. Indianapolis, IN: Que. [This is the second edition of a widely cited critique of information overload by a well-known architect and student of the work of Louis Kahn.]

Zimmerli, W. 1986. "Who is to blame for data pollution?" In C. Mitcham and A. Huning, eds., *Philosophy and Technology II: Information Technology and Computers in Theory and Practice*. Boston: Reidel. [One of 20 original papers from a conference, introduced by an overview of "Information technology and computers as themes in the philosophy of technology," and followed by an annotated bibliography on philosophical studies of information technology and computers.]

Chapter 26

Computational Modeling as a Philosophical Methodology

Patrick Grim

Since the sixties, computational modeling has become increasingly important in both the physical and the social sciences, particularly in physics, theoretical biology, sociology, and economics. Since the eighties, philosophers too have begun to apply computational modeling to questions in logic, epistemology, philosophy of science, philosophy of mind, philosophy of language, philosophy of biology, ethics, and social and political philosophy. This chapter analyzes a selection of interesting examples in some of these areas.

Computer Models in the Sciences and Philosophy: Benefits and Limitations

What qualifies as a computer model or a computer simulation has itself been subject to philosophical scrutiny, but without clear consensus. In a classic statement, T. Naylor (1966) defines computer simulation as

> a numerical technique for conducting experiments on a digital computer which involves certain types of mathematical and logical models that describe the behavior of . . . systems over extended periods of time. (p. 3)

On the other hand, Fritz Rohrlich (1991) and Paul Humphreys (1991), among others, emphasize the importance of tractability as a motivation for computer modeling. Humphreys builds that feature into his working definition:

> A computer simulation is any computer-implemented method for exploring the properties of mathematical models where analytic methods are unavailable. (1991: 501)

Most authors have not attempted strict definition, conceding that the notion of a "model" is vague and may even have several distinct senses (Fetzer 1999). Still, several important features of models are repeatedly emphasized in the literature: (1) models occupy a conceptual role somewhere between empirical data and traditional theory; (2) modeling represents a wide variety of techniques, rather than a single tool; and (3) model construction itself is part of the "art" of science (see especially Rohrlich 1991). There is also general agreement on reasons for welcoming

computer modeling in particular: (1) increased mathematical tractability, particularly in understanding complex and dynamic processes over time; (2) a methodologically important vividness or graphic immediacy that is often characteristic of computer models; and (3) the possibility of computational "experiment." This last feature is clear to anyone who has worked with computer models, and is noted in almost every outline of computational modeling since Naylor (1966), quoted above.

A number of authors portray computer experimentation in general as a technological extension of an ancient tradition of thought experiment. It is this "experimental" aspect of computational modeling that has been seen as a particularly important addition to philosophical methodology. Kyburg (1998: 37) speaks of a "kind of philosophical laboratory or testing ground." Grim, Mar, and St. Denis (1998: 10) speak of "an important new *environment* for philosophical research," and Bynum and Moor (1998: 6) speak of computing as "a medium in which to model philosophical theories and positions":

> Computing and related concepts significantly enhance philosophy by providing a kind of intellectual clay that philosophers can mold and shape and study. Through computing, abstract ideas – which philosophers like to manipulate – can be instantiated and investigated. There is nothing wrong with good armchair reflection . . . But armchair reflection has its limitations. (1998: 2–3)

The exploration of abstract philosophical ideas by means of computer models offers a number of major benefits. One benefit is an astounding increase in manageable complexity. Although philosophers have long appealed to thought experiments, practical necessity has limited these to our individual human powers of calculation. As Bynum and Moor note, "armchair recursion doesn't recur very many times." With computer models, on the other hand, the computational ceiling is lifted on philosophical imagination. Complex interactions that previously could only be vaguely guessed at can now be calculated with ease, and consequences of such interactions can be revealed with a depth previously impossible. The development of complex systems over time could hardly have been envisaged at all before the computer, but has now become a topic of philoso-phical thought experiment in a wide range of areas.

Another benefit of computer modeling is that its methodological demands work as a counterforce against philosophical vices of imprecision, vagueness, and obscurity. The environment of computer modeling enforces "unflinchingly and without compromise, the central philosophical desideratum of clarity: one is forced to construct theory in the form of fully explicit models, so detailed and complete that they can be *programmed*" (Grim, Mar, & St. Denis 1998: 10). John Pollock has emphasized that one constraint imposed by computer modeling is simply that the theory at issue must actually work the way it is supposed to. "As mundane as this constraint may seem, I am convinced that most epistemological theories fail to satisfy it." The fact that computer modeling imposes demands of precision and detail "can have a very therapeutic effect on a profession that is overly fond of handwaving" (Pollock 1998: 34).

Another benefit of a computational environment is the prospect of exploring possible variations on theory. With a computer model in place, variations on the model are generally easy. One can explore consequences of epistemological, biological, or social theories in slightly different environments or with slightly different parameters. The result is that the theory with which one begins may be replaced by a variation that appears more promising in action.

As has always been true of the interplay between technology and pure science, it is also possible for the application of philosophical ideas in computational models at the bottom to suggest new and intriguing philosophical ideas at the top. It was computer work in graphing the semantics of infinite-valued paradox, for example, that suggested the proof for a theorem on formal undefinability of chaos (Grim, Mar, & St. Denis 1998). Models developed in order to tackle old problems may open up new territory for philosophical exploration as well. A central question in Hobbes is how cooperation can emerge in a society of self-serving egoists. Game-theoretic attempts to answer that question have been

developed and expanded in computer models (see Chapter 22, GAME THEORY). Those models have in turn raised further questions regarding evolution and ethics, rationality and justice, and the unpredictability of social behavior in even our simplest models (Axelrod 1984, Danielson 1992, Skyrms 1996, Grim, Mar, & St. Denis 1998).

Some benefits of computational modeling are evident even in aspects for which it is occasionally criticized. Formal models in general and philosophical computer models in particular are bound to be abstract. The high level of abstraction, however, can be seen not as a weakness but as an indication of potential power and promise. Distinct phenomena that appear in quite different contexts – biology and economics, for example, or logic and physics – may nonetheless have a similar structure or exhibit a similar dynamics. The search for patterns that hold across different disciplines has characterized new fields such as chaos theory and artificial life (see Chapter 15), and promises to be an area in which philosophical computer modeling could flourish as well. As Bedau (1998: 135) remarks,

> By abstracting away from the details of chaotic systems (such as ecologies, turbulent fluid flow, and economic markets), chaos science seeks fundamental properties that unify and explain a diverse range of chaotic systems. Similarly, by abstracting away from the details of life-like systems (such as ecologies, immune systems, and autonomously evolving social groups) and synthesizing these processes in artificial media, typically computers, the field of artificial life seeks to understand the essential processes shared by broad classes of life-like systems.

One promise of philosophical computer modeling is highly abstract crossdisciplinary work of precisely this type.

Models in the physical and particularly the social sciences are also sometimes met with the objection that they are *mere* models: that the phenomena being modeled are complex in ways to which a simple model could not possibly do justice. The same is to be expected as an occasional response to philosophical computer modeling. In reply, it must simply be admitted that all models have major limitations built in. That is part of what makes them models. Models, like abstract laws, prove useful in both explanation and exploration precisely *because* they are simpler than the full phenomenon, and thus easier to handle and track. We need simple models because we need a simpler way of understanding complex phenomena, and because we need to separate what is important in what happens from the distracting but unimportant details.

One can then argue that neither the abstract level nor the simplicity of models requires further defense. With computational modeling as with any methodology, however, there are some real intellectual dangers. New methodologies always offer new ways of approaching particular *kinds* of questions. There is thus always a temptation to phrase questions in only those ways that the new method can handle, or to ask only those questions that the new method can easily address. We may end up considering only those versions of a theory that can be readily modeled, for example, or attending only to those types of theory that can be modeled at all. The only known cure for such a danger is to be aware of it. Although we do not want to ignore promising new techniques, we must be aware that they will inevitably come with limitations. For any promising new tool, there will always be further questions, equally deep and serious, for which it may *not* be the best approach.

Computational models carry a more specific danger as well. As models increase in sophistication in a particular tradition of model-building, they are inevitably built out of their simpler predecessors. Computational models often incorporate earlier passages of code wholesale. Thus, if early models in a tradition carry an inessential feature or an unexamined assumption, that feature or assumption is likely to remain, unquestioned and uncriticized, throughout later work as well. One of the earliest models to be applied to questions in economics, for example, was built using a quite particular balance of potential gains and losses: the gains and losses characteristic of the Prisoner's Dilemma, discussed further below. There is now extensive research in theoretical biology, evolutionary psychology, and philosophy using the descendants of that original model. But it is almost never asked whether the particular gains and losses built into the model reflect

realistic assumptions for the specific applications at issue.

It must finally be admitted that individual models can certainly fail. Within both philosophy and the sciences a simple model may turn out to be too simple, or may be simple in the wrong ways. It may abstract not from accidental features but from fundamentally essential aspects of what is being modeled. The possibility always remains that a model captures too few aspects of the phenomenon, or the wrong ones. That models can go wrong in these ways is grounds for criticism of individual models, of course, but it constitutes no objection against a methodology of model-building in general. When a model falls short it quite generally suggests a better model.

The following sections emphasize several areas of current exploration in philosophical computer modeling. Evident in much of this work is a strong interdisciplinary or cross-disciplinary tendency. Modeling techniques developed primarily within physics have been applied within logic (Grim, Mar, & St. Denis 1998); techniques developed within computer science have been brought together with traditions in economics and sociology and applied to questions of ethics and social-political philosophy (Skyrms 1996, Danielson 1992, Grim, Mar, & St. Denis 1998). The comments of Clark Glymour and his collaborators with regard to "android epistemology" can be applied to philosophical computer modeling more generally:

> The force of this idea can be seen in the way in which it violates all kinds of traditional disciplinary boundaries in science, bringing together engineering and the life sciences, placing mathematical linguistics in the heart of electrical engineering and requiring moral philosophers to understand computation theory. University deans, forced to work within the old hierarchies, weep with frustration, and the work often has to be done in new "interdisciplinary" – and often undisciplined – research centers and institutes which escape the old categories. We live in interesting times. (Ford, Glymour, & Hayes 1995: xii)

The rest of this chapter is devoted to some key examples of philosophical modeling being done in these interesting times.

Logic

At a fundamental level, computers are logical machines: their basic operations can be outlined in terms of standard logical connectives. This immediately suggests the possibility of turning to computers as tools for extending work in traditional logic. One might expect intensive work on theorem-proving programs, for example, and indeed research of this type has an impressive history (see especially the bibliography on logic and automated theorem proving at <www.cs.cmu.edu/afs/cs/project/pal/bibs/Logictext.bib>). What is interesting, however, is that computers have also played a key role in the development of *non*traditional logical models.

Suppose we start from a set of premises P_1, and add a few more to create a larger set of premises P_2. In traditional logic, anything we could have deduced from premises P_1 will also be deducible from the inclusive set P_2: classical logics are *monotonic*. Much of everyday reasoning, however, seems to be nonmonotonic. If I am told that Tweety is a bird, I conclude that Tweety can fly. If given more information – that Tweety is a bird and a penguin, for example – I may withdraw that commitment. The need to handle "defeasible" or nonmonotonic reasoning of this kind quickly became evident in attempts at modeling patterns of reasoning in artificial intelligence, and the development of rival approaches is active and ongoing (Pollock 1998, Kyburg 1998, Gabbay, Hogger, & Robinson 1994).

The fact that computer models can offer vivid images of abstract phenomena is exploited for logical purposes in Grim, Mar, and St. Denis 1998. Here simple considerations from truth-table semantics are extended to construct "value solids," which portray in spatial terms the combinatory operation of connectives within particular systems (figure 26.1). Spatial representation of logical properties can make formal relations immediately apparent: in figure 26.1, for example, the duality of conjunction and disjunction is reflected in the fact that the value solid for "and" could be inverted and inserted into that for "or." Modeling of this sort has also produced some surprises, such as the persistent reappearance of the fractal Sierpinski gasket in a range of

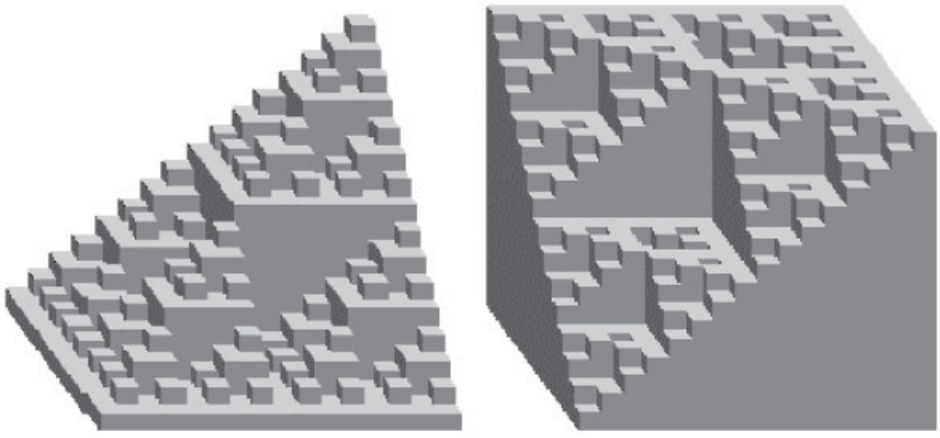

Figure 26.1: Value solids for AND (left) and OR (right)

value solids and the possibility of generating value solids by cellular-automata-like rules.

The fact that computer models can capture complex dynamics is used by Grim (1993) and Grim, Mar, and St. Denis (1998) in work on self-reference and paradox in infinite-valued logics. Informally presented, the Liar sentence ("This sentence is false") seems to produce an alternation between truth and falsity: "if it's true, since it says it's false, it must be false . . . but if it's false, and it says it's false, it must be true . . ." That dynamics is modeled in the first frame of figure 26.2. The authors consider relatives of the Liar within infinite-valued logics, such as the Chaotic Liar – "this sentence is as true as you think it is false" – which has a dynamics fully chaotic in the mathematical sense, illustrated for slight differences in initial estimated value in the second frame of figure 26.2.

Computer-modeling prospects for new approaches to logic are developed in a different way in John Barwise and Jon Etchemendy's *Hyperproof* (1994). In previous work, Barwise and Etchemendy had developed *Tarski's World* as a visual aid for teaching quantificational logic, with *Turing's World* as a similarly visual introduction to Turing machines. Sensitized from that experience to the power and ubiquity of visualization in reasoning, Barwise and Etchemendy's goal in *Hyperproof* is to expand logic beyond its current ties to sentential syntax to a formalization of information-processing that can exploit various forms of representation, diagrammatic as well as sentential (figure 26.3). Their purposes go far beyond the pedagogical. Traditional logic, as they portray it, has concentrated on only a "narrow slice" of a broader realm of valid information-extraction. "In the long run, logic must come to grips with how people use a multitude of representations in rigorous ways. This will force us to extend and enrich the traditional notions of syntax, semantics, logical consequence and proof . . . In the process, what seemed like a finished success story in philosophical and mathematical analysis will be refashioned in exciting new ways" (Barwise & Etchemendy 1998: 112).

Epistemology

One of the goals of epistemology is to understand how we come to know. It is related forms of engineering – epistemology "from the design stance," "epistemological engineering," or "android epistemology" – that have produced exciting research programs in philosophical computer modeling. It is clear that all of these will overlap with the tasks of artificial intelligence more generally (see Chapter 9, The Philosophy of AI and its Critique). Here as elsewhere in computer modeling, interdisciplinary collaboration is the rule rather than the exception.

John Pollock's OSCAR Project takes as its objective "the construction of a general theory of rationality and its computer-implementation

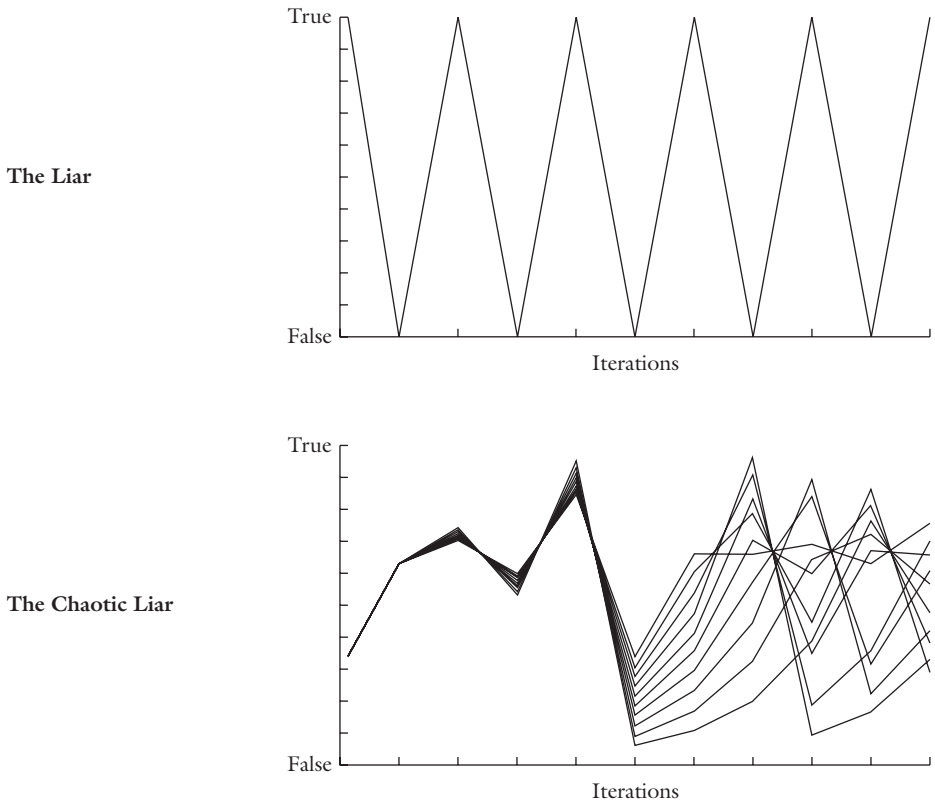

Figure 26.2: Dynamics of the Liar and the Chaotic Liar over progressive iterations. The Chaotic Liar is shown for small differences in initial values.

in an artificial rational agent" (Pollock 1998: 17). The general epistemic task is taken to be that of systematically updating a set of beliefs towards those which are warranted, and OSCAR is designed to exploit techniques of both defeasible and deductive logic toward that end. Pollock places particular emphasis on control of an epistemic system in terms of interests and goals. In epistemology, as elsewhere in philosophy, one of the benefits of computer modeling is the necessity of making assumptions explicit in a way that also exposes them to fruitful criticism. The fact that the program avoids Bayesian probability theory is taken as a point in favor of OSCAR's rationality by Pollock, but forms the basis for a number of criticisms in Kyburg (1998). The project is described in Pollock (1989) and (1995), and a current version of OSCAR is downloadable from <www.u.arizona.edu/~pollock/>.

Paul Thagard has used computational modeling in pursuing a wide range of issues in philosophy of science and epistemology more generally (see Chapter 23, COMPUTING IN THE PHILOSOPHY OF SCIENCE). ECHO was developed as a connectionist computational model of explanatory coherence, and Thagard has applied it to a number of examples from the history of science and in critique of other approaches (Thagard 1992, Eliasmith & Thagard 1997). He has also attempted to model major aspects of analogical thinking using the programs ARCS and ACME (Holyoak & Thagard, 1997). The deeply interdisciplinary character of much of his research is particularly clear in the work on induction (Holland, Holyoak, Nisbett, & Thagard 1987), in which Thagard works with two psychologists and a computer scientist in attempting to construct a computational model

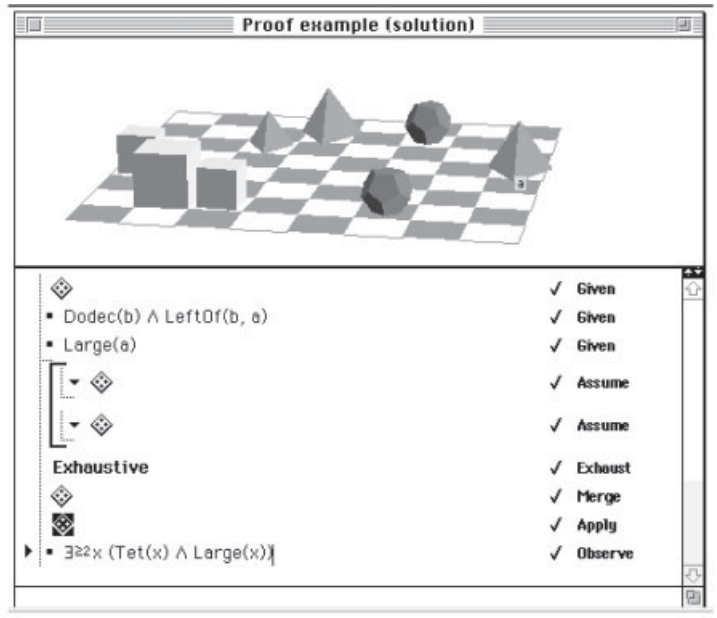

Figure 26.3: A diagrammatic proof from Barwise and Etchemendy's *Hyperproof*

of scientific reasoning which avoids traditional philosophical problems.

The TETRAD project should also be mentioned as a rigorous computational attempt to treat epistemological issues regarding causal inference and probability with an eye to issues in experimental methodology (Spirtes, Glymour, & Scheines 2001). A very different computational approach explores the complex epistemological dynamics of competing information. Some of the information received by an epistemic agent is about the accuracy or reliability of information sources. In simple models, this can both produce a variety of epistemic chaos and suggest maps for the management of epistemic chaos (Grim, Mar, & St. Denis 1998).

Philosophy of Mind, Philosophy of Language, and Philosophy of Biology

It could be argued that the entire field of AI qualifies as computational modeling in philosophy of mind. Traditional debates regarding innatism and empiricism, the character and limits of the human mind, and even freedom and consciousness are now debated with illustrations drawn from competing computational architectures. Work by John Searle, Daniel Dennett, Jerry Fodor, and David Chalmers features prominently in the philosophical debate, if not so prominently in the details of computational modeling. Research by the Churchlands is noteworthy for framing contemporary debates in terms of current resources in neurophysiology and computer science, with an emphasis on neural networks as models for both the power and the peculiar inscrutability of the workings of the human mind (P. S. Churchland & Sejnowski 1993, P. M. Churchland 1995). A variety of attempts to approach issues in philosophy of mind using the tools of dynamical systems theory are represented in Robert Port and Timothy van Gelder (1997).

Some attempts have recently been made to apply tools of computational modeling to issues in the philosophy of language. Structural mapping approaches to analogy and ambiguity appear in Holyoak and Thagard 1997. They form the core of a program designed to generate and interpret metaphors in Steinhart and Kittay 1994 and

Steinhart 1995. Computer modeling can also be expected to play an important role in the inevitable conflict between Chomskian models of linguistic representation and language learning and alternative connectionist proposals (McClelland, St. John, & Taraban 1992). A game-theoretic attempt to understand linguistic convention initiated in Lewis's *Convention* (1969) is further developed with the tools of replicator dynamics in Skyrms's *Evolution of the Social Contract* (1996). In this same tradition, simple computational environments which show emergence of coordinated signaling behavior are offered as models for a theory of meaning as use in Grim, Kokalis, Tafti, & Kilb 2000 and 2002, and Grim, St. Denis, & Kokalis 2003.

Much as AI can be seen as a computational version of philosophy of mind, Artificial Life (ALife) can be seen as a computational version of philosophy of biology (see Chapter 15). Mark Bedau uses ALife to illustrate both his theory of life as supple adaptation (1996) and his consideration of emergent phenomena in biology (1997). In time, it seems inevitable that tools from AI and ALife will be brought together to answer questions in philosophy of mind and philosophy of biology. Some simple mathematical models in this direction are offered in Peter Godfrey-Smith 1996.

Ethics, Social and Political Philosophy

Computational modeling and ethics might seem an unlikely combination, but these are linked in an intriguing interdisciplinary history. Game theory (see Chapter 22) was developed by von Neumann and Morgenstern (1944) as an attempt at a mathematical theory applicable to economics and political strategy. The Prisoner's Dilemma, a two-person game that seems to capture a basic tension between individual and collective advantage, quickly became a paradigm for work in aspects of economics, theoretical sociology, and eventually theoretical biology. Played in terms of "cooperations" and "defections" on each side, the Prisoner's Dilemma has been referred to as the *e. coli* of social psychology.

In 1980, political scientist Robert Axelrod invited experts in game theory from various fields to submit programs for a Computer Prisoner's Dilemma Tournament (Axelrod 1984). Submitted strategies played 200 games against all other strategies, themselves, and a strategy that chose responses at random. The winner of that tournament was a strategy called "Tit for Tat" (TFT). Cooperate with TFT and it will cooperate with you on the next round. Defect against TFT and it will defect against you. The fact that such a cooperative strategy triumphed in the first tournament was a surprise. Its continued success in later tournaments, where its reputation clearly made it the strategy to beat, was a further surprise. Axelrod and Hamilton (1981) replaced the tournament competition with an "ecological model," which employs the replicator dynamics of theoretical biology: more successful strategies "breed" to occupy a larger percentage of the population. Here too TFT triumphs. The affinity of these results with the Hobbesian question of how social cooperation can grow in a community of self-serving egoists is striking, and the contemporary models inevitably drew the attention of philosophers. Could the triumph of something that looked like altruism in this formal model be telling us something about the dynamics or nature of social cooperation and ethics? Brian Skyrms has pursued this cluster of philosophical questions using the tools of computer modeling, adding also techniques drawn from dynamical systems or chaos theory. "Using these tools of evolutionary dynamics, we can now study aspects of the social contract from a fresh perspective" (1996: p. x). In *Evolution and the Social Contract*, Skyrms shows that an evolutionary model using replicator dynamics goes beyond rational decision theory in producing particular principles of fair division and a "law of mutual aid." The potentially chaotic dynamics of a bargaining game is illustrated in figure 26.4.

Grim, Mar, & St. Denis (1998) emphasize spatialized models of emerging cooperation. Figure 26.5, for example, shows a spatialized conquest by TFT in a field of 8 simple strategies. Nowak and Sigmund (1992) showed that a greater level of cooperation ("generous TFT," which forgives defection against it in 1/3 of all cases) triumphs in versions of Axelrod and

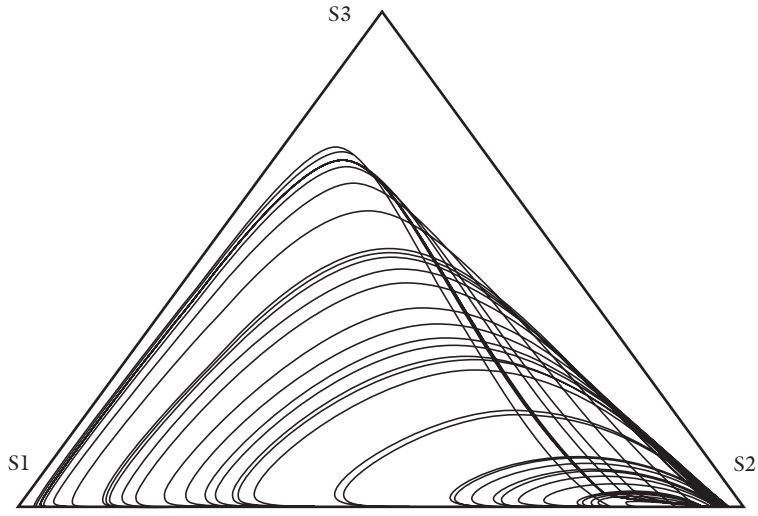

Figure 26.4: A chaotic attractor in game-theoretical dynamics (from Skyrms 1997)

Hamilton's model that are arguably more realistic in incorporating stochastic or probabilistic imperfections. Grim, Mar and St. Denis show that spatialization of such models results in an even higher level of cooperation. They also consider a version of the spatialized model with "fuzzy" values of intermediate cooperation and defection, show a formal undecidability result for the Spatialized Prisoner's Dilemma, and take some first steps toward applying the model to questions of racial discrimination.

Peter Danielson characterizes his work in this tradition as "artificial morality," intended to combine game theory and artificial intelligence in the development of "instrumental contractarianism" as an ethical theory. Building on the work of David Gauthier's *Morals by Agreement* (1986) and constructing a range of PROLOG models, Danielson's attempt is to show at least that it can be rational to be moral. Danielson seems more willing than other researchers in the area, however, to use modeling as an argument for something more: for a version of ethical naturalism in which morality simply *is* that strategy that proves successful. In some forms of such a view, such as Michael Ruse's Darwinian naturalism, the conclusion is that morality is something other than what we have thought it to be: "Morality is no more than a collective illusion fobbed off on us by our genes for reproductive ends" (Ruse 1991: 506). The use of computational models to demonstrate naturalistic conclusions of this sort is contested in Grim, Mar, and St. Denis 1998. They offer as a counterexample the success of certain discriminatory strategies, which play TFT with others of their own color but always defect against outsiders. Strategic success cannot simply be identified with morality, they argue, since discriminatory strategies are clearly successful in such models, but it is clear that analogous racial discrimination is morally wrong. How social strategies may develop or propagate is one question; whether they should be judged as genuinely ethical is another.

Powerful new tools are now available for further research in this general tradition. TIERRA and the later AVIDA are ALife platforms that may be customized to pursue questions in both philosophy of biology and social and political philosophy. TIERRA is available by anonymous ftp from <ftp://alife.santafe.edu>. A good introduction to AVIDA, packaged with the software, is Adami, *Introduction to Artificial Life* (1998). SWARM is a powerful general platform for agent-based modeling, developed by members of the Santa Fe Institute as a way of offering researchers in various fields a powerful "lingua franca" for computational experimentation. The program can be downloaded at <www.swarm.org>. A good introduction to this platform is Luna and

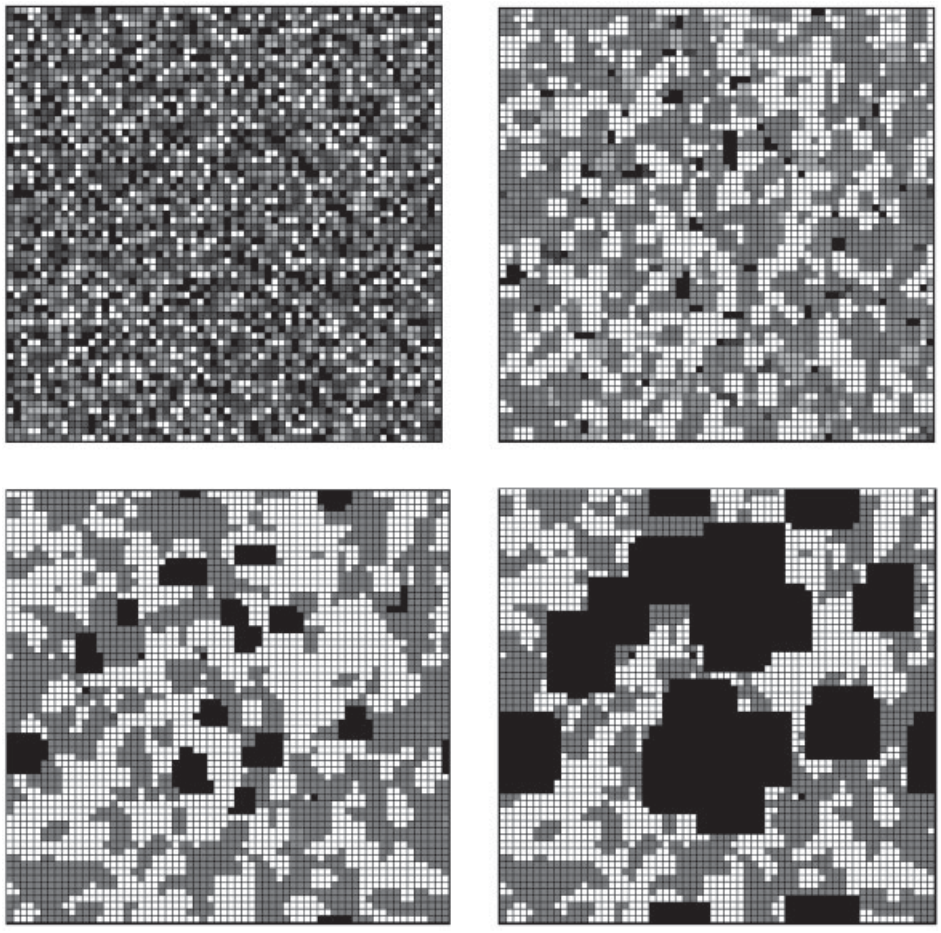

Figure 26.5: Progressive conquest by Tit for Tat, shown in black, in an array of 8 simple strategies. TFT eventually conquers the entire field.

Stefansson (2000), which offers a SWARM instantiation for the Spatialized Prisoner's Dilemma as one of its opening examples.

Conclusion

Computational modeling offers not a single systematized method but an enormous toolbox of potential models and techniques. A number of basic modeling tools – the tools of dynamical systems, neural nets, and cellular automata, for example – have found application across the physical and social sciences. The result has been active model-borrowing between different research programs and a flourishing interdisciplinary awareness. The application of computational modeling to philosophical questions has just begun, and first returns are promising in a number of areas.

It must be recognized that novel techniques carry some intellectual risks. Powerful new methods inevitably draw attention to those questions, or those forms of questions, for which the methods hold out the most promise. In philosophy as elsewhere it must be remembered that the new techniques should take their place among a range of traditional tools for approaching a range of perennial questions. It is also important

to beware of fixing on a few simple models too early. Continuing development of new variations is important in order to avoid narrow assumptions and to facilitate wide exploration.

Progress in philosophical computer modeling is already proceeding so swiftly that any overview is bound to be partial and incomplete. In this chapter, the attempt has been to emphasize the general promise of such an approach by highlighting a few examples of intriguing and noteworthy current work.

Acknowledgments

I am grateful to Jason Alexander, Mark Bedau, Terrell Bynum, John Etchemendy, James Fetzer, Henry Kyburg, Paul Humphreys, James Moor, Brian Skyrms, and Eric Steinhart for helpful input, and to members of the Group for Logic & Formal Semantics and Luciano Floridi for careful comments on earlier drafts.

References

Adami, C. 1998. *Introduction to Artificial Life*. New York: Springer-Verlag. [A good introduction to the AVIDA program for artificial life. Advanced undergraduate and above.]

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books. [A major source for game-theoretic modeling regarding competition and cooperation, and a good introduction for any interested reader.]

—— and Hamilton, W. 1981. "The evolution of cooperation." *Science* 211: 1390–6. [Application of replicator dynamics to earlier game-theoretic work regarding competition and cooperation. Technical but clear.]

Barwise, J. and Etchemendy, J. 1994. *Hyperproof*. Stanford, CA: CSLI and Cambridge University Press. [A system of computer-enhanced logic. Advanced undergraduate and above.]

—— and ——. 1998. "Computers, visualization, and the nature of reasoning." In T. Bynum and J. Moor, eds., *The Digital Phoenix*. Oxford: Blackwell, pp. 93–116. [Working reflections on computer prospects for new approaches to logic. Undergraduate and above.]

Bedau, M. 1996. "The nature of life." In M. Boden, ed., *The Philosophy of Artificial Life*. Oxford: Oxford University Press, pp. 331–57. [Bedau's theory of life as supple adaptation illustrated in terms of artificial life. Suitable for any interested reader.]

——. 1997. "Weak emergence." In J. Tomberlin, ed., *Philosophical Perspectives: Mind, Causation, and World* (vol. 11). Oxford: Blackwell, pp. 375–99. [One of the best philosophical treatments of the notion of emergent phenomena. Requires some philosophical background.]

——. 1998. "Philosophical content and method of artificial life." In Bynum & Moor 1998: 37–47. [A good general discussion of artificial life as computational philosophy of biology. Undergraduate and above.]

——. 1999. "Can unrealistic computer models illuminate theoretical biology?" In A. Wu, ed., *Proceedings of the 1999 Genetic and Evolutionary Computation Conference Workshop Program*, pp. 20–3. Available on-line at <www.reed.edu/~mab/papers.htm>. [An argument for the virtues of abstract modeling. Suitable for any interested reader.]

Burkholder, L. 1992. *Philosophy and the Computer*. Oxford: Westview Press. [A valuable anthology of computer-assisted philosophical work earlier than Bynum & Moor 1998. Different contributions at different levels of difficulty.]

Bynum, T. and Moor, J., eds. 1998. *The Digital Phoenix: How Computers are Changing Philosophy*. Oxford: Blackwell. [An important anthology of work in computer modeling, from which several other references are drawn. A good introduction for anyone.]

Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul*. Cambridge, MA: MIT Press. [A philosophically informed review of current work in neural nets, including also neurophysiology. Suitable for any interested reader.]

Churchland, P. S. and Sejnowski, T. 1993. *The Computational Brain*. Cambridge, MA: MIT Press. [A philosophically informed review of current work in neurophysiology, including also work on neural nets. Occasionally technical but clearly written.]

Danielson, P. 1992. *Artificial Morality: Virtuous Robots for Virtual Games*. New York: Routledge. [Uses formal game-theoretic modeling to argue for a form of ethical naturalism. Undergraduate and above.]

Eliasmith, C. and Thagard, P. 1997. "Waves, particles, and explanatory coherence." *British Journal for the Philosophy of Science* 48: 1–19. [An application of ECHO to the history of the wave-particle debate in physics. Some background in physics required.]

Fetzer, J. 1999. "The role of models in computer science." *The Monist* 82: 20–36. [A critical examination of various approaches to models. Advanced undergraduate and above.]

Ford, K., Glymour, C., and Hayes, P. 1995. *Android Epistemology*. Menlo Park, CA: AAAI Press/MIT Press. [An anthology of philosophical reflections on epistemological modeling. Undergraduate and above.]

Gabbay, D., Hogger, C., and Robinson, J. 1994. *Handbook of Logic in Artificial Intelligence and Logic Programming, vol. 3: Nonmonotonic Reasoning and Uncertain Reasoning*. Oxford: Clarendon Press. [An overview of formal work in defeasible and nonmonotonic logics. Very technical, graduate and above.]

Gauthier, D. 1986. *Morals by Agreement*. Oxford: Oxford University Press. [A contractarian model important for Danielson's "artificial morality" models. Undergraduate and above.]

Godfrey-Smith, P. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press. [Mathematical models relevant to both philosophy of mind and philosophy of biology. Advanced undergraduate and above.]

Grim, P. 1993. "Self-reference and chaos in fuzzy logic." *IEEE Transactions on Fuzzy Systems* 1: 237–53. [Dynamical models for the behavior of self-referential sentences in fuzzy logic. Sometimes technical, graduate and above.]

——, Kokalis, T., Tafti, A., and Kilb, N. 2000. "Evolution of communication in perfect and imperfect worlds." *World Futures* 56: 179–97. [A spatialized model in which simple patterns of communication emerge as behavioral coordination within a community. Advanced undergraduate and above.]

——, ——, ——, and ——. 2002. "Evolution of communication with a spatialized genetic algorithm." *Evolution of Communication* 3: 105–34. [A genetic algorithm model showing emergence of simple patterns of communication in order to capture food and avoid predation. Advanced undergraduate and above.]

——, Mar, G., and St. Denis, P. 1998. *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. Cambridge, MA: MIT Press. [A rich sampler of philosophical computer modeling by a single research group, including a CD-ROM with animations and all source code. Undergraduate and above.]

——, St. Denis, P. and Kokalis, T. 2003. "Learning to communicate: the emergence of signaling in spatialized arrays of neural nets." Group for Logic & Formal Semantics, Dept. of Philosophy, SUNY Stony Brook, Research Report no. 01-01, forthcoming in *Adaptive Behavior*. [A neural net model showing emergence of simple patterns of communication in order to capture food and avoid predation. Advanced undergraduate level and above.]

Holland, J., Holyoak, K., Nisbett, R., and Thagard, P. 1987. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press. [A better philosophy of science through computer modeling. Graduate and above.]

Holyoak, K. and Thagard, P. 1997. "The analogical mind." *American Psychologist* 52: 35–44. [An attempt to model reasoning by analogy. Advanced undergraduate and above.]

Humphreys, P. 1991. "Computer simulations." In A. Fine, M. Forbes, and L. Wessels, eds., *PSA 1990*, vol. 2: 497–506. East Lansing, MI: Philosophy of Science Association. [A philosophical examination of computer models in the sciences. Occasionally technical, advanced undergraduate and above.]

Kyburg, H. 1998. "Epistemology and computing." In Bynum & Moor 1998: 37–47. [Includes criticism of Pollock's OSCAR project. Suitable for any interested reader.]

Luna, F. and Stefansson, B., eds. 2000. *Economic Simulations in Swarm: Agent-based Modeling and Object Oriented Programming*. Boston: Kluwer Academic. [A good introduction to a new platform for social, and biological experimentation developed by the Santa Fe Institute and downloadable at <www.swarm.org>. Different elements accessible at advanced undergraduate and graduate level.]

McClelland, J. L., St. John, M., and Taraban, R. 1992. "Sentence comprehension: a parallel distributed processing approach." In L. Burkholder, ed., *Philosophy and the Computer*. Oxford: Westview, pp. 34–56. [An opening salvo in the growing controversy between Chomskian linguistics and connectionist models. Suitable for any interested reader.]

Naylor, T. 1966. *Computer Simulation Techniques.* New York: Wiley. [An early text, now of mostly historical interest. Advanced undergraduate.]

Nowak, M. and Sigmund, K. 1992. "Tit for tat in heterogeneous populations." *Nature* 355: 250–52. [Introducing stochastic imperfection into game-theoretic models of cooperation. The result: greater generosity. Advanced undergraduate and above.

Pollock, J. 1989. *How to Build a Person: A Prolegomenon.* Cambridge, MA: MIT Press. [Problems faced and lessons drawn from the OSCAR project. Undergraduate and above.]

——. 1995. *Cognitive Carpentry.* Cambridge, MA: MIT Press. [Problems faced and lessons drawn from the OSCAR project. Undergraduate and above.]

——. 1998. "Procedural epistemology." In Bynum & Moor 1998: 17–36. [An outline of the OSCAR project. Undergraduate and above.]

Port, R. and van Gelder, T., eds. 1997. *Mind as Motion: Explorations in the Dynamics ff Cognition.* Cambridge, MA: MIT Press. [An anthology of work modeling psychological processes in terms of dynamical systems theory. Different contributions at different levels.]

Rohrlich, P. 1991. "Computer simulations in the physical sciences." In A. Fine, M. Forbes, and L. Wessels, eds., *PSA 1990*, vol. 2: 507–18. East Lansing, MI: Philosophy of Science Association. [A philosophical examination of computer models in the sciences. Graduate and above.]

Ruse, M. 1991. "The significance of evolution." In P. Singer, ed., *A Companion to Ethics.* Oxford: Blackwell, pp. 500–510. [A particularly strong version of ethical naturalism. Suitable for any interested reader.]

Skyrms, B. 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press. [A seminal source of formal game-theoretic modeling relevant to questions in ethics and social and political philosophy. Undergraduate and above.]

——. 1997. "Chaos and the explanatory significance of equilibrium: strange attractors in evolutionary game dynamics." In C. Bicchieri, R. Jeffrey, and B. Skyrms, eds., *The Dynamics of Norms.* Cambridge: Cambridge University Press. [Game-theoretical dynamics with an emphasis on chaos. Sometimes technical, graduate and above.]

Spirtes, P., Glymour, C., and Scheines, R. 2001. *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press. [A sophisticated attempt to model causation and prediction well enough to improve research design. Technical, graduate and above.]

Steinhart, E. 1995. "NETMET: a program for generating and interpreting metaphors." *Computers and Humanities* 28: 383–92. [An attempt at programming semantic "fields." Undergraduate and above.]

—— and Kittay, E. 1994 "Generating metaphors from networks." In J. Hintikka, ed., *Approaches to Metaphor.* Dordrecht: Kluwer Academic, pp. 41–94. [An attempt at programming semantic "fields." Undergraduate and above.]

Thagard, P. 1992. *Conceptual Revolutions.* Princeton: Princeton University Press. [The ECHO program applied to examples from the history of science. Undergraduate and above.]

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press. [The seminal work on game theory. Graduate and above.]

Index

Note: Abbreviations used in the index: AI = artificial intelligence; ALife = artificial life; CMC = computer-mediated communication; GDI = General Definition of Information; HCI = human–computer interaction; IT = information technology; MTC = mathematical theory of communication; SDI = Special Definition of Information; VR = virtual reality.