Reasoning about Knowledge

Ronald Fagin
Joseph Y. Halpern
Yoram Moses
Moshe Y. Vardi

The MIT Press
Cambridge, Massachusetts
London, England

Reasoning about Knowledge

Ronald Fagin
Joseph Y. Halpern
Yoram Moses
Moshe Y. Vardi

The MIT Press
Cambridge, Massachusetts
London, England

First MIT Press paperback edition, 2003

This book was set in Times Roman and MathTime by Windfall Software (using LaTeX) and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Reasoning about knowledge / Ronald Fagin . . . [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-262-06162-7 (hardcover: alk. paper), 0-262-56200-6 (pb)
1. Knowledge, Theory of. 2. Agent (Philosophy) 3. Reasoning.
I. Fagin, Ronald.
BD181.R38 1995
153.4'—dc20 94-36477
CIP

10 9 8 7 6 5 4 3 2

Reasoning about Knowledge

To Susan, who is as happy and amazed as I am that The Book is finally completed; to Josh, Tim, and Teddy, who are impressed that their father is an Author; and to the memories of my mother Maxine, who gave me a love of learning, and of my father George, who would have been proud.

R. F.

To Gale, for putting up with this over the years; to David and Sara, for sometimes letting Daddy do his work; and to my mother Eva, to whom I can finally say "It's done!"

J. Y. H.

To my father Shimon, to Yael, Lilach and Eyal, and to the memory of my mother Ala and my brother Amir. With Love.

Y. M.

To Pam, who listened for years to my promises that the book is 90% done; to Aaron, who, I hope, will read this book; to my parents, Ziporah and Pinkhas, who taught me to think; and to my grandparents, who perished in the Holocaust.

הביטו וראו, אם־יש מכאוב כמכאובי.

"Behold and see, if there be any sorrow like unto my sorrow."

M. Y. V.

Preface to the Hardcover Edition

As its title suggests, this book investigates reasoning about knowledge, in particular, reasoning about the knowledge of agents who reason about the world and each other's knowledge. This is the type of reasoning one often sees in puzzles or Sherlock Holmes mysteries, where we might have reasoning such as this:

> If Alice knew that Bob knew that Charlie was wearing a red shirt, then Alice would have known that Bob would have known that Charlie couldn't have been in the pantry at midnight. But Alice didn't know this ...

As we shall see, this type of reasoning is also important in a surprising number of other contexts. Researchers in a wide variety of disciplines, from philosophy to economics to cryptography, have all found that issues involving agents reasoning about other agents' knowledge are of great relevance to them. We attempt to provide here a framework for understanding and analyzing reasoning about knowledge that is intuitive, mathematically well founded, useful in practice, and widely applicable.

The book is almost completely self-contained. We do expect the reader to be familiar with propositional logic; a nodding acquaintance with distributed systems may be helpful to appreciate some of our examples, but it is not essential. Our hope is that the book will be accessible to readers from a number of different disciplines, including computer science, artificial intelligence, philosophy, and game theory. While proofs of important theorems are included, the non-mathematically-oriented reader should be able to skip them, while still following the main thrust of the book.

We have tried to make the book modular, so that, whenever possible, separate chapters can be read independently. At the end of Chapter 1 there is a brief overview of the book and a table of dependencies. Much of this material was taught a number of times by the second author in one-quarter courses at Stanford University and

by the third author in one-semester courses at the Weizmann Institute of Science. Suggestions for subsets of material that can be covered can also be found at the end of Chapter 1.

Many of the details that are not covered in the main part of the text of each chapter are relegated to the exercises. As well, the exercises cover material somewhat tangential—but still of interest!—to the main thrust of the chapter. We recommend that the reader at least look over all the exercises in each chapter. Far better, of course, would be to do them all (or at least a reasonable subset). Problems that are somewhat more difficult are marked with $*$, and even more difficult problems are marked with $**$.

Each chapter ends with a section of notes. These notes provide references to the material covered in each chapter (as well as the theorems that are stated but not proved) and, occasionally, more details on some points not covered in the chapter. The references appearing in the notes are to the latest version of the material we could find. In many cases, earlier versions appeared in conference proceedings. The dates of the references that appear in the notes therefore do not provide a chronological account of the contributions to the field. While we attempt to provide reasonably extensive coverage of the literature in these notes, the field is too large for our coverage to be complete. We apologize for the inadvertent omission of relevant references.

The book concludes with a bibliography, a symbol index, and an index.

Many people helped us in many ways in the preparation of this book, and we are thankful to all of them. Daphne Koller deserves a very special note of thanks. She did a superb job of proofreading the almost-final draft of the book. Besides catching many typographical errors, she gave us numerous suggestions on improving the presentation in every chapter. We are very grateful to her. We would also like to thank Johan van Benthem, Adam Grove, Vassos Hadzilacos, Lane Hemaspaandra and the students of CS 487 at the University of Rochester, Wil Janssen, Hector Levesque, Murray Mazer, Ron van der Meyden, Jan Pachl, Karen Rudie, Ambuj Singh, Elias Thijsse, Mark Tuttle, and Lenore Zuck, for their useful comments and criticisms; Johan van Benthem, Brian Chellas, David Makinson, and Krister Segerberg for their help in tracking down the history of modal logic; and T. C. Chen and Brian Coan for pointing out the quotations at the beginning of Chapters 2 and 3, respectively. Finally, the second and third authors would like to thank the students of CS 356 (at Stanford in the years 1984–1989, 1991–1992, and 1994), CS 2422S (at Toronto in 1990) and the course on Knowledge Theory (at the Weizmann Institute of Science in the years 1987–1995), who kept finding typographical errors and suggesting improvements to

the text (and wondering if the book would ever be completed), especially Gidi Avrahami, Ronen Brafman, Ed Brink, Alex Bronstein, Isis Caulder, Steve Cummings, John DiMarco, Kathleen Fisher, Steve Friedland, Tom Henzinger, David Karger, Steve Ketchpel, Orit Kislev, Christine Knight, Ronny Kohavi, Rick Kunin, Sherry Listgarten, Carlos Mendioroz, Andres Modet, Shahid Mujtaba, Gal Nachum, Leo Novik, Raymond Pang, Barney Pell, Sonne Preminger, Derek Proudian, Omer Reingold, Tselly Regev, Gil Roth, Steve Souder, Limor Tirosh-Pundak-Mintz, Maurits van der Veen, Orli Waarts, Scott Walker, and Liz Wolf.

Finally, we wish to thank the institutions that supported this work for many years; the work of the first, second, and fourth authors was done at the IBM Almaden Research Center, and the work of the third author was done at the Weizmann Institute of Science, and while on sabbatical at the Oxford University Computing Laboratory. The work of the third author was supported in part by a Sir Charles Clore Post-Doctoral Fellowship, by an Alon Fellowship, and by a Helen and Milton A. Kimmelman Career Development Chair.

Preface to the Paperback Edition

Relatively few changes have been made for this edition of the book. For the most part, this involved correcting typos and minor errors and updating references. Perhaps the most significant change involved moving material from Chapter 7 on a notion called "nonexcluding contexts" back to Chapter 5, and reworking it. This material is now used in Chapter 6 to refine the analysis of the interaction between common knowledge and agreement protocols.

The effect of teaching a number of classes using the hardcover edition of the book can be seen in this edition. The second author would like to thank the students of CS 676 (at Cornell in the years 1996, 1998, and 2000) for their comments and suggestions, especially Wei Chen, Francis Chu, David Kempe, Yoram Minsky, Nat Miller, and Suman Ganguli. The third author would like to thank the students of the course "Knowledge and Games in Distributed Systems" (at the Technion EE dept. in the years 1998, 2000, and 2002) for their comments and suggestions, especially Tomer Koll, Liane Levin, and Alex Sprintson. We would also like to thank Jelle Gerbrandy for pointing a minor bug in Chapter 3, and Rohit Parikh for pointing out minor bugs in Chapters 1 and 2.

The second and third authors changed institutions between the hardcover and paperback editions. The fourth author moved shortly before the hardcover edition appeared. The second author is now at Cornell University, the third author is at the Technion, and the fourth author is at Rice University. We would like to thank these institutions for their support of the work on the paperback edition.

Contents

Preface to the Hardcover Edition xi

Preface to the Paperback Edition xv

1 Introduction and Overview 1
- 1.1 The Muddy Children Puzzle 4
- 1.2 An Overview of the Book 8

2 A Model for Knowledge 15
- 2.1 The Possible-Worlds Model 15
- 2.2 Adding Common Knowledge and Distributed Knowledge 23
- 2.3 The Muddy Children Revisited 25
- 2.4 The Properties of Knowledge 31
- 2.5 An Event-Based Approach 38

3 Completeness and Complexity 49
- 3.1 Completeness Results 51
- 3.2 Decidability 65
- 3.3 Incorporating Common Knowledge 70
- 3.4 Incorporating Distributed Knowledge 73
- 3.5 The Complexity of the Validity Problem 74
- 3.6 *NP*-Completeness Results for S5 and KD45 78
- 3.7 The First-Order Logic of Knowledge 80
  - 3.7.1 First-Order Logic 81
  - 3.7.2 First-Order Modal Logic 84
  - 3.7.3 Assumptions on Domains 86

3.7.4 Properties of Knowledge in Relational Kripke Structures . . 87

4 Knowledge in Multi-Agent Systems 109
4.1 Runs and Systems . . . 109
4.2 Incorporating Knowledge . . . 116
4.3 Incorporating Time . . . 121
4.4 Examples of Systems . . . 123
4.4.1 Knowledge Bases . . . 123
4.4.2 Game Trees . . . 131
4.4.3 Synchronous Systems . . . 135
4.4.4 Perfect Recall . . . 136
4.4.5 Message-Passing Systems . . . 138
4.4.6 Asynchronous Message-Passing Systems . . . 141
4.5 Knowledge Gain in A.M.P. Systems . . . 145

5 Protocols and Programs 163
5.1 Actions . . . 163
5.2 Protocols and Contexts . . . 167
5.3 Programs . . . 180
5.4 Specifications . . . 183

6 Common Knowledge and Agreement 189
6.1 Coordinated Attack . . . 190
6.2 Agreeing to Disagree . . . 199
6.3 Simultaneous Byzantine Agreement . . . 206
6.4 Nonrigid Sets and Common Knowledge . . . 213
6.5 Attaining SBA . . . 218
6.6 Attaining Common Knowledge . . . 224
6.6.1 Clean Rounds . . . 225
6.6.2 Waste . . . 226
6.6.3 Computing Common Knowledge . . . 230
6.7 Detailed Proofs . . . 232

7 Knowledge-Based Programming 253
7.1 Knowledge-Based Programs . . . 253
7.2 Getting Unique Representations . . . 259
7.3 Knowledge Bases Revisited . . . 271

7.4 A Knowledge-Based Program for SBA 276
7.5 Strong Correctness 281
7.6 The Sequence-Transmission Problem 283
7.7 Proving Strong Correctness of ST 290

8 Evolving Knowledge **303**
8.1 Properties of Knowledge and Time 303
8.2 Synchrony and Perfect Recall 307
8.3 Knowledge and Time in A.M.P. Systems 311
8.4 Knowledge and Time in $\mathcal{I}_n^{oa}(\Phi)$ 313
8.5 A Closer Look at Axiom $\mathrm{OA}_{n,\Phi}$ 318

9 Logical Omniscience **333**
9.1 Logical Omniscience 334
9.2 Explicit Representation of Knowledge 337
9.2.1 The Syntactic Approach 338
9.2.2 The Semantic Approach 340
9.2.3 Discussion 345
9.3 Nonstandard Logic 346
9.3.1 Nonstandard Structures 346
9.3.2 Strong Implication 350
9.3.3 A Payoff: Querying Knowledge Bases 354
9.3.4 Discussion 357
9.4 Impossible Worlds 357
9.5 Awareness 362
9.6 Local Reasoning 368
9.7 Concluding Remarks 373

10 Knowledge and Computation **391**
10.1 Knowledge and Action Revisited 391
10.2 Algorithmic Knowledge 394
10.2.1 Algorithmic Knowledge Systems 394
10.2.2 Properties of Algorithmic Knowledge 398
10.3 Examples 399
10.4 Algorithmic Knowledge Programs 402
10.4.1 Algorithmic Knowledge Programming 403

10.4.2 Algorithmic Knowledge and Complexity 405
10.4.3 Implementing Knowledge-Based Programs 408

11 Common Knowledge Revisited **415**
11.1 Common Knowledge as a Conjunction 416
11.2 Common Knowledge and Simultaneity 419
11.2.1 Common Knowledge and Uncertainty 419
11.2.2 Simultaneous Events 421
11.3 Temporal Imprecision 425
11.4 The Granularity of Time 428
11.5 Common Knowledge as a Fixed Point 433
11.5.1 Fixed Points 433
11.5.2 Downward Continuity and Infinite Conjunctions 440
11.6 Approximations of Common Knowledge 443
11.6.1 ε- and Eventual Common Knowledge 443
11.6.2 Applications to Coordinated Attack 447
11.6.3 Timestamped Common Knowledge 451
11.6.4 Other Approximations of Common Knowledge 453
11.7 Discussion 454

Bibliography **463**

Symbol Index **489**

Index **493**

Chapter 1

Introduction and Overview

An investment in knowledge pays the best interest.

Benjamin Franklin, *Poor Richard's Almanac*, c. 1750

Epistemology, the study of knowledge, has a long and honorable tradition in philosophy, starting with the early Greek philosophers. Questions such as "What do we know?" "What can be known?" and "What does it mean to say that someone knows something?" have been much discussed in the philosophical literature. The idea of a formal logical analysis of reasoning about knowledge is somewhat more recent, but goes back at least to von Wright's work in the early 1950's. The first book-length treatment of *epistemic logic*—the logic of knowledge—is Hintikka's seminal work *Knowledge and Belief*, which appeared in 1962. The 1960's saw a flourishing of interest in this area in the philosophy community. The major interest was in trying to capture the inherent properties of knowledge. Axioms for knowledge were suggested, attacked, and defended.

More recently, researchers in such diverse fields as economics, linguistics, AI (artificial intelligence), and theoretical computer science have become interested in reasoning about knowledge. While, of course, some of the issues that concerned the philosophers have been of interest to these researchers as well, the focus of attention has shifted. For one thing, there are pragmatic concerns about the relationship between knowledge and action. What does a robot need to know in order to open a safe, and how does it know whether it knows enough to open it? At what point does an economic agent know enough to stop gathering information and make a decision? When should a database answer "I don't know" to a query? There are also concerns

about the complexity of computing knowledge, a notion we can now quantify better thanks to advances in theoretical computer science. Finally, and perhaps of most interest to us here, is the emphasis on considering situations involving the knowledge of a group of agents, rather than that of just a single agent.

When trying to understand and analyze the properties of knowledge, philosophers tended to consider only the single-agent case. But the heart of any analysis of a conversation, a bargaining session, or a protocol run by processes in a distributed system is the interaction between agents. The focus of this book is on understanding the process of reasoning about knowledge in a group and using this understanding to help us analyze complicated systems. Although the reader will not go far wrong if he or she thinks of a "group" as being a group of people, it is useful to allow a more general notion of "group," as we shall see in our applications. Our agents may be negotiators in a bargaining situation, communicating robots, or even components such as wires or message buffers in a complicated computer system. It may seem strange to think of wires as agents who know facts; however, as we shall see, it is useful to ascribe knowledge even to wires.

An agent in a group must take into account not only facts that are true about the world, but also the knowledge of other agents in the group. For example, in a bargaining situation, the seller of a car must consider what the potential buyer knows about the car's value. The buyer must also consider what the seller knows about what the buyer knows about the car's value, and so on. Such reasoning can get rather convoluted. Most people quickly lose the thread of such nested sentences as "Dean doesn't know whether Nixon knows that Dean knows that Nixon knows that McCord burgled O'Brien's office at Watergate." But this is precisely the type of reasoning that is needed when analyzing the knowledge of agents in a group.

A number of states of knowledge arise naturally in a multi-agent situation that do not arise in the one-agent case. We are often interested in situations in which *everyone* in the group knows a fact. For example, a society certainly wants all drivers to know that a red light means "stop" and a green light means "go." Suppose we assume that every driver in the society knows this fact and follows the rules. Will a driver then feel safe? The answer is no, unless she also knows that everyone else knows and is following the rules. For otherwise, a driver may consider it possible that, although she knows the rules, some other driver does not, and that driver may run a red light.

Even the state of knowledge in which everyone knows that everyone knows is not enough for a number of applications. In some cases we also need to consider the state in which simultaneously everyone knows a fact φ, everyone knows that everyone

knows φ, everyone knows that everyone knows that everyone knows φ, and so on. In this case we say that the group has *common knowledge* of φ. This key notion was first studied by the philosopher David Lewis in the context of conventions. Lewis pointed out that in order for something to be a convention, it must in fact be common knowledge among the members of a group. (For example, the convention that green means "go" and red means "stop" is presumably common knowledge among the drivers in our society.) John McCarthy, in the context of studying common-sense reasoning, characterized common knowledge as what "any fool" knows; "any fool" knows what is commonly known by all members of a society.

Common knowledge also arises in discourse understanding. Suppose that Ann asks Bob "What did you think of the movie?" referring to a showing of *Monkey Business* they have just seen. Not only must Ann and Bob both know that "the movie" refers to *Monkey Business*, but Ann must know that Bob knows (so that she can be sure that Bob will give a reasonable answer to her question), Bob must know that Ann knows that Bob knows (so that Bob knows that Ann will respond appropriately to his answer), and so on. In fact, by a closer analysis of this situation, it can be shown that there must be common knowledge of what movie is meant in order for Bob to answer the question appropriately.

Finally, common knowledge also turns out to be a prerequisite for achieving agreement. This is precisely what makes it such a crucial notion in the analysis of interacting groups of agents.

At the other end of the spectrum from common knowledge is distributed knowledge. A group has distributed knowledge of a fact φ if the knowledge of φ is distributed among its members, so that by pooling their knowledge together the members of the group can deduce φ, even though it may be the case that no member of the group individually knows φ. For example, if Alice knows that Bob is in love with either Carol or Susan, and Charlie knows that Bob is not in love with Carol, then together Alice and Charlie have distributed knowledge of the fact that Bob is in love with Susan, although neither Alice nor Charlie individually has this knowledge. While common knowledge can be viewed as what "any fool" knows, distributed knowledge can be viewed as what a "wise man"—one who has complete knowledge of what each member of the group knows—would know.

Common knowledge and distributed knowledge are useful tools in helping us understand and analyze complicated situations involving groups of agents. The puzzle described in the next section gives us one example.

1.1 The Muddy Children Puzzle

Reasoning about the knowledge of a group can involve subtle distinctions between a number of states of knowledge. A good example of the subtleties that can arise is given by the "muddy children" puzzle, which is a variant of the well known "wise men" or "cheating wives" puzzles.

> Imagine n children playing together. The mother of these children has told them that if they get dirty there will be severe consequences. So, of course, each child wants to keep clean, but each would love to see the others get dirty. Now it happens during their play that some of the children, say k of them, get mud on their foreheads. Each can see the mud on others but not on his own forehead. So, of course, no one says a thing. Along comes the father, who says, "At least one of you has mud on your forehead," thus expressing a fact known to each of them before he spoke (if $k > 1$). The father then asks the following question, over and over: "Does any of you know whether you have mud on your own forehead?" Assuming that all the children are perceptive, intelligent, truthful, and that they answer simultaneously, what will happen?
>
> There is a "proof" that the first $k - 1$ times he asks the question, they will all say "No," but then the k^{th} time the children with muddy foreheads will all answer "Yes."
>
> The "proof" is by induction on k. For $k = 1$ the result is obvious: the one child with a muddy forehead sees that no one else is muddy. Since he knows that there is at least one child with a muddy forehead, he concludes that he must be the one. Now suppose $k = 2$. So there are just two muddy children, a and b. Each answers "No" the first time, because of the mud on the other. But, when b says "No," a realizes that he must be muddy, for otherwise b would have known the mud was on his forehead and answered "Yes" the first time. Thus a answers "Yes" the second time. But b goes through the same reasoning. Now suppose $k = 3$; so there are three muddy children, a, b, c. Child a argues as follows. Assume that I do not have mud on my forehead. Then, by the $k = 2$ case, both b and c will answer "Yes" the second time. When they do not, he realizes that the assumption was false, that he is muddy, and so will answer "Yes" on the third question. Similarly for b and c.
>
> The argument in the general case proceeds along identical lines.