Reasoning about Knowledge

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 978-0-262-06162-9 (hc.: alk. paper), 978-0-262-56200-3 (pb.:alk. paper)
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

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.

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 488

Index 491

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.