Reasoning about Knowledge

Ronald Fagin

Joseph Y. Halpern

Yoram Moses

Moshe Y. Vardi

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

1

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 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 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 kth 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 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.

Let us denote the fact "at least one child has a muddy forehead" by p. Notice that if $k > 1$, i.e., more than one child has a muddy forehead, then every child can see at least one muddy forehead, and the children initially all know p. Thus, it would seem that the father does not provide the children with any new information, and so

he should not need to tell them that p holds when $k > 1$. But this is false! In fact, as we now show, if the father does not announce p, the muddy children are never able to conclude that their foreheads are muddy.

Here is a sketch of the proof: We prove by induction on q that, no matter what the situation is, i.e., no matter how many children have a muddy forehead, all the children answer "No" to the father's first q questions. Clearly, no matter which children have mud on their foreheads, all the children answer "No" to the father's first question, since a child cannot tell apart a situation where he has mud on his forehead from one that is identical in all respects except that he does not have a muddy forehead. The inductive step is similar: By the inductive hypothesis, the children answer "No" to the father's first q questions. Thus, when the father asks his question for the $(q + 1)^{st}$ time, child i still cannot tell apart a situation where he has mud on his forehead from one that is identical in all respects except that he does not have a muddy forehead, since by the induction hypothesis, the children will answer "No" to the father's first q questions whether or not child i has a muddy forehead. Thus, again, he does not know whether his own forehead is muddy.

So, by announcing something that the children all know, the father somehow manages to give the children useful information! How can this be? Exactly what *is* the role of the father's statement? Of course, the father's statement did enable us to do the base case of the induction in the proof, but this does not seem to be a terribly satisfactory answer. It certainly does not explain what information the children gained as a result of the father's statement.

We can answer these questions by using the notion of common knowledge described in the previous section. Let us consider the case of two muddy children in more detail. It is certainly true that before the father speaks, everyone knows p. But it is not the case that everyone knows that everyone knows p. If Alice and Bob are the only children with muddy foreheads, then before the father speaks, Alice considers it possible that she does not have mud on her forehead, in which case Bob does not see anyone with a muddy forehead and so does not know p. After the father speaks, Alice does know that Bob knows p. After Bob answers "No" to the father's first question, Alice uses her knowledge of the fact that Bob knows p to deduce that her own forehead is muddy. (Note that if Bob did not know p, then Bob would have said "No" the first time even if Alice's forehead were clean.)

We have just seen that if there are only two muddy children, then it is not the case that everyone knows that everyone knows p before the father speaks. However, if there are three muddy children, then it *is* the case that everyone knows that everyone knows p before the father speaks. If Alice, Bob, and Charlie have muddy foreheads, then Alice knows that Bob can see Charlie's muddy forehead, Bob knows that Charlie

can see Alice's muddy forehead, etc. It is not the case, however, that everyone knows that everyone knows that everyone knows p before the father speaks. In general, if we let $E^k p$ represent the fact that everyone knows that everyone knows ... (k times) p, and let Cp represent the fact that p is common knowledge, then we leave it to the reader to check that if exactly k children have muddy foreheads, then $E^{k-1} p$ holds before the father speaks, but $E^k p$ does not. It turns out that when there are k muddy children, $E^k p$ suffices to ensure that the children with muddy foreheads will be able to figure it out, while $E^{k-1} p$ does not. The father's statement actually converts the children's state of knowledge from $E^{k-1} p$ to Cp. With this extra knowledge, they can deduce whether their foreheads are muddy.

The careful reader will have noticed that we made a number of implicit assumptions in the preceding discussion over and above the assumption made in the story that "the children are perceptive, intelligent, and truthful." Suppose again that Alice and Bob are the only children with muddy foreheads. It is crucial that both Alice and Bob *know* that the children are intelligent, perceptive, and truthful. For example, if Alice does not know that Bob is telling the truth when he answers "No" to the father's first question, then she cannot answer "Yes" to the second question (even if Bob is in fact telling the truth). Similarly, Bob must know that Alice is telling the truth. Besides its being known that each child is intelligent, perceptive, and truthful, we must also assume that each child knows that the others can see, that they all hear the father, that the father is truthful, and that the children can do all the deductions necessary to answer the father's questions.

Actually, even stronger assumptions need to be made. If there are k children with muddy foreheads, it must be the case that everyone knows that everyone knows ... ($k - 1$ times) that the children all have the appropriate attributes (they are perceptive, intelligent, all hear the father, etc.). For example, if there are three muddy children and Alice considers it possible that Bob considers it possible that Charlie might not have heard the father's statement, then she cannot say "Yes" to the father's third question (even if Charlie in fact did hear the father's statement and Bob knows this). In fact, it seems reasonable to assume that all these attributes are common knowledge, and, indeed, this assumption seems to be made by most people on hearing the story.

To summarize, it seems that the role of the father's statement was to give the children common knowledge of p (the fact that at least one child has a muddy forehead), but the reasoning done by the children assumes that a great deal of common knowledge already existed in the group. How does this common knowledge arise? Even if we ignore the problem of how facts like "all the children can see" and "all

the children are truthful" become common knowledge, there is still the issue of how the father's statement makes p common knowledge.

Note that it is not quite correct to say that p becomes common knowledge because all the children hear the father. Suppose that the father had taken each child aside individually (without the others noticing) and said "At least one of you has mud on your forehead." The children would probably have thought it a bit strange for him to be telling them a fact that they already knew. It is easy to see that p would not become common knowledge in this setting.

Given this example, one might think that the common knowledge arose because all the children *knew* that they all heard the father. Even this is not enough. To see this, suppose the children do not trust each other, and each child has secretly placed a miniature microphone on all the other children. (Imagine that the children spent the previous summer at a CIA training camp.) Again the father takes each child aside individually and says "At least one of you has a muddy forehead " In this case, thanks to the hidden microphones, all the children know that each child has heard the father, but they still do not have common knowledge.

A little more reflection might convince the reader that the common knowledge arose here because of the *public* nature of the father's announcement. Roughly speaking, the father's public announcement of p puts the children in a special situation, one with the property that all the children know both that p is true and that they are in this situation. We shall show that under such circumstances p is common knowledge. Note that the common knowledge does not arise because the children somehow deduce each of the facts $E^k p$ one by one. (If this were the case, then arguably it would take an infinite amount of time to attain common knowledge.) Rather, the common knowledge arises all at once, as a result of the children being in such a special situation. We return to this point in later chapters.

1.2 An Overview of the Book

The preceding discussion should convince the reader that the subtleties of reasoning about knowledge demand a careful formal analysis. In Chapter 2, we introduce a simple, yet quite powerful, formal semantic model for knowledge, and a language for reasoning about knowledge. The basic idea underlying the model is that of *possible worlds*. The intuition is that if an agent does not have complete knowledge about the world, she will consider a number of worlds possible. These are her candidates for the way the world actually is. The agent is said to *know* a fact φ if φ holds at all the worlds that the agent considers to be possible. Using this semantic model allows

us to clarify many of the subtleties of the muddy children puzzle in quite an elegant way. The analysis shows how the children's state of knowledge changes with each response to the father's questions, and why, if there are k muddy children altogether, it is only after the kth question that the children with muddy foreheads can deduce this fact.

We should emphasize here that we do not feel that the semantic model we present in the next chapter is the unique "right" model of knowledge. We spend some time discussing the properties of knowledge in this model. A number of philosophers have presented cogent arguments showing that some of these properties are "wrong." Our concerns in this book are more pragmatic than those of the philosophers. We do not believe that there is a "right" model of knowledge. Different notions of knowledge are appropriate for different applications. The model we present in the next chapter is appropriate for analyzing the muddy children puzzle and for many other applications, even if it is not appropriate for every application. One of our goals in this book is to show how the properties of "knowledge" vary with the application.

In Chapter 3, we give a complete characterization of the properties of knowledge in the possible-worlds model. We describe two approaches to this characterization. The first approach is *proof-theoretic*: we show that all the properties of knowledge can be formally proved from the properties discussed in Chapter 2. The second approach is *algorithmic*: we study algorithms that can determine whether a given property holds under our definition of knowledge, and consider the computational complexity of doing this.

One of the major applications we have in mind is using knowledge to analyze *multi-agent systems*, be they systems of interacting agents or systems of computers in a network. In Chapter 4 we show how we can use our semantic model for knowledge to *ascribe* knowledge to agents in a multi-agent system. The reason that we use the word "ascribe" here is that the notion of knowledge we use in the context of multi-agent systems can be viewed as an *external* notion of knowledge. There is no notion of the agent computing his knowledge, and no requirement that the agent be able to answer questions based on his knowledge. While this may seem to be an unusual way of defining knowledge, we shall argue that it does capture one common usage of the word "know." Moreover, we give examples that show its utility in analyzing multi-agent systems.

In Chapter 5 we extend the model of Chapter 4 to consider *actions*, *protocols*, and *programs*. This allows us to analyze more carefully how changes come about in multi-agent systems. We also define the notion of a *specification* and consider what it means for a protocol or program to satisfy a specification.

In Chapter 6 we show how useful a knowledge-based analysis of systems can be. Our focus in this chapter is common knowledge, and we show how fundamental it is in various contexts. In particular, we show that it is a prerequisite for agreement and simultaneous coordinated action.

In Chapter 7 we extend our notions of programs to consider *knowledge-based* programs, which allow explicit tests for knowledge. Knowledge-based programs can be viewed as giving us a high-level language in which to program or specify a system. We give a number of examples showing the usefulness of thinking and programming at the knowledge level.

In Chapter 8 we consider the properties of knowledge and time, focusing on how knowledge evolves over time in multi-agent systems. We show that small changes in the assumptions we make about the interactions between knowledge and time in a system can have quite subtle and powerful effects on the properties of knowledge.

As we show in Chapter 2, one property that seems to be an inherent part of the possible-worlds model of knowledge is that agents are *logically omniscient*. Roughly speaking, this means they know all tautologies and all logical consequences of their knowledge. In the case of the muddy children puzzle we explicitly make the assumption that each child can do all the reasoning required to solve the puzzle. While this property may be reasonable for some applications, it certainly is not reasonable in general. After all, we cannot really hope to build logically omniscient robots. In Chapter 9 we describe several approaches for constructing abstract models that do not have the logical omniscience property.

As we have already discussed, our notion of knowledge in multi-agent systems is best understood as an external one, ascribed by, say, the system designer to the agents. We do not assume that the agents compute their knowledge in any way, nor do we assume that they can necessarily answer questions based on their knowledge. In a number of applications that we are interested in, agents need to *act* on their knowledge. In such applications, external knowledge is insufficient; an agent that has to act on her knowledge has to be able to compute this knowledge. The topic of knowledge and computation is the subject of Chapter 10.

In Chapter 11, we return to the topic of common knowledge. We suggested in the previous section that common knowledge arose in the muddy children puzzle because of the public nature of the father's announcement. In many practical settings such a public announcement, whose contents are understood simultaneously by many agents, is impossible to achieve. We show that, in a precise sense, common knowledge cannot be attained in these settings. This puts us in a somewhat paradoxical situation, in that we claim both that common knowledge is a prerequisite for agreement and coordinated action and that it cannot be attained. We examine this
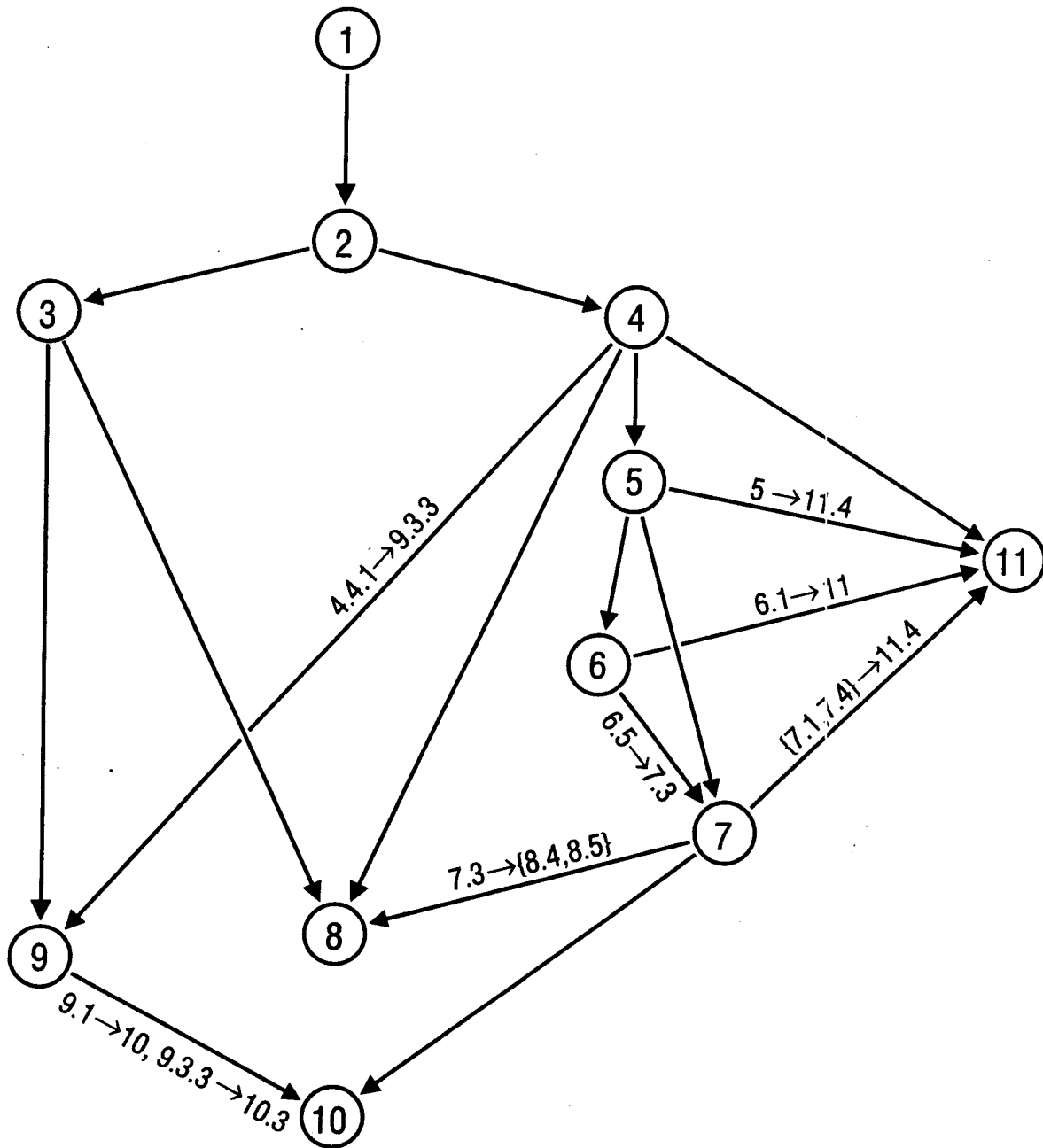
Figure 1.1   Dependency diagram

paradox in Chapter 11 and suggest a number of weaker notions that can be attained in practice and that are often sufficient for acting in the real world.

Although a considerable amount of the material in this book is based on previously published work, a number of elements are new. These include much of the material in Chapters 5, 7, 10, and some of Chapter 11. Specifically, the notions of contexts and programs in Chapter 5, and of knowledge-based programs and their implementation in Chapter 7, are new. Moreover, they play a significant role in the way we model and analyze knowledge and action in multi-agent systems.

We have tried as much as possible to write the book in a modular way, so that material in the later chapters can be read without having to read all the preceding chapters. Figure 1.1 describes the dependencies between chapters. An arrow from one chapter to another indicates that it is necessary to read (at least part of) the first chapter in order to understand (at least part of) the second. We have labeled the arrow if it is not necessary to read all of the first chapter to understand all of the second. For example, the label $9.1 \rightarrow 10$, $9.3.3 \rightarrow 10.3$ on the arrow from Chapter 9 to Chapter 10 indicates that the only sections in Chapter 9 on which Chapter 10 depends are 9.1 and 9.3.3 and, moreover, the only section in Chapter 10 that depends on Section 9.3.3 is Section 10.3. Similarly, the label $5 \rightarrow 11.4$ on the arrow from Chapter 5 to Chapter 11 indicates that Section 11.4 is the only section in Chapter 11 that depends on Chapter 5, but it depends on the whole chapter.

Certain material can be skipped without losing a broad overview of the area. In particular, this is the case for Sections 3.3, 3.4, 4.5, 6.7, and 7.7. The second author covered a substantial portion of the remaining material (moving at quite a rapid pace) in a one-quarter course at Stanford University. A course designed to focus on the application of our approach to distributed systems could cover Chapters 1, 2, 4, 5, 6, 7, 10, and 11. Each chapter ends with exercises and bibliographic notes; these could be useful in a course based on this book. As we mentioned in the preface, we strongly recommend that the reader at least look over the exercises.

Exercises

1.1 The *aces and eights* game is a simple game that involves some sophisticated reasoning about knowledge. It is played with a deck consisting of just four aces and four eights. There are three players. Six cards are dealt out, two to each player. The remaining two cards are left face down. Without looking at the cards, each of the players raises them up to his or her forehead, so that the other two players can see

them but he or she cannot. Then all of the players take turns trying to determine which cards they're holding (they do not have to name the suits). If a player does not know which cards he or she is holding, the player must say so. Suppose Alice, Bob, and you are playing the game. Of course, it is common knowledge that none of you would ever lie, and that you are all perfect reasoners.

(a) In the first game, Alice, who goes first, holds two aces, and Bob, who goes second, holds two eights. Both Alice and Bob say that they cannot determine what cards they are holding. What cards are you holding? (Hint: consider what would have happened if you held two aces or two eights.)

(b) In the second game, you go first. Alice, who goes second, holds two eights. Bob, who goes third, holds an ace and an eight. No one is able to determine what he or she holds at his or her first turn. What do you hold? (Hint: by using part (a), consider what would have happened if you held two aces.)

(c) In the third game, you go second. Alice, who goes first, holds an ace and an eight. Bob, who goes third, also holds an ace and an eight. No one is able to determine what he or she holds at his or her first turn; Alice cannot determine her cards at her second turn either. What do you hold?

* **1.2** Show that in the aces and eights game of Exercise 1.1, someone will always be able to determine what cards he or she holds. Then show that there exists a situation where only one of the players will be able to determine what cards he or she holds, and the other two will never be able to determine what cards they hold, no matter how many rounds are played.

1.3 The *wise men puzzle* is a well-known variant of the muddy children puzzle. The standard version of the story goes as follows: There are three wise men. It is common knowledge that there are three red hats and two white hats. The king puts a hat on the head of each of the three wise men, and asks them (sequentially) if they know the color of the hat on their head. The first wise man says that he does not know; the second wise man says that he does not know; then the third wise man says that he knows.

(a) What color is the third wise man's hat?

(b) We have implicitly assumed in the story that the wise men can all see. Suppose we assume instead that the third wise man is blind and that it is common knowledge that the first two wise men can see. Can the third wise man still figure out the color of his hat?

Notes

The idea of a formal logical analysis of reasoning about knowledge seems to have first been raised by von Wright [1951]. As we mentioned in the text, Hintikka [1962] gave the first book-length treatment of epistemic logic. Lenzen [1978] gives an overview of the work in epistemic logic done in the 1960s and 1970s. He brings out the arguments for and against various axioms of knowledge. The most famous of these arguments is due to Gettier [1963], who argued against the classical interpretation of knowledge as true, justified belief; his work inspired many others. Gettier's arguments and some of the subsequent papers are discussed in detail by Lenzen [1978]. For recent reviews of the subject, see the papers by Halpern [1986, 1987, 1993a], by Meyer, van der Hoek, and Vreeswijk [1991a, 1991b], by Moses [1992], and by Parikh [1990].

As we mentioned, the original work on common knowledge was done by Lewis [1969] in the context of studying conventions. Although McCarthy's notion of what "any fool" knows goes back to roughly 1970, it first appears in a published paper in [McCarthy, Sato, Hayashi, and Igarishi 1979]. The notion of knowledge and common knowledge has also been of great interest to economists and game theorists, ever since the seminal paper by Aumann [1976]. Knowledge and common knowledge were first applied to multi-agent systems by Halpern and Moses [1990] and by Lehmann [1984]. The need for common knowledge in understanding a statement such as "What did you think of the movie?" is discussed by Clark and Marshall [1981]; a dissenting view is offered by Perrault and Cohen [1981]. Clark and Marshall also present an example of nested knowledge based on the Watergate scandal, mentioning Dean and Nixon. The notion of distributed knowledge was introduced by Halpern and Moses [1990]. They initially called it *implicit knowledge*, and the term "distributed knowledge" was suggested by Jan Pachl.

The muddy children puzzle is a variant of the "unfaithful wives" puzzle discussed by Gamow and Stern [1958]. Gardner [1984] also presents a variant of the puzzle, and a number of variants of the puzzle are discussed by Moses, Dolev, and Halpern [1986]. The version given here is taken almost verbatim from [Barwise 1981]. The aces and eights puzzle in Exercise 1.1 is taken from [Carver 1989]. Another related puzzle is the so-called "Conway paradox", which was first discussed by Conway, Paterson, and Moscow [1977], and later by Gardner [1977]. It was analyzed in an epistemic framework by van Emde Boas, Groenendijk, and Stokhof [1980]. An extension of this puzzle was considered by Parikh [1992].

See the bibliographic notes in later chapters for more references on the specific subjects discussed in these chapters.

Chapter 2

A Model for Knowledge

Chuangtse and Hueitse had strolled onto the bridge over the Hao, when the former observed, "See how the small fish are darting about! That is the happiness of the fish." "You are not a fish yourself," said Hueitse. "How can you know the happiness of the fish?" "And you not being I," retorted Chuangtse, "how can you know that I do not know?"

Chuangtse, c. 300 B.C.

2.1   The Possible-Worlds Model

As we said in Chapter 1, our framework for modeling knowledge is based on *possible worlds*. The intuitive idea behind the possible-worlds model is that besides the true state of affairs, there are a number of other possible states of affairs, or "worlds." Given his current information, an agent may not be able to tell which of a number of possible worlds describes the actual state of affairs. An agent is then said to *know* a fact φ if φ is true at all the worlds he considers possible (given his current information). For example, agent 1 may be walking on the streets in San Francisco on a sunny day but may have no information at all about the weather in London. Thus, in all the worlds that the agent considers possible, it is sunny in San Francisco. (We are implicitly assuming here that the agent does not consider it possible that he is hallucinating and in fact it is raining heavily in San Francisco.) On the other hand, since the agent has no information about the weather in London, there are worlds he considers possible in which it is sunny in London, and others in which it is raining in London. Thus, this agent knows that it is sunny in San Francisco,

15

but he does not know whether it is sunny in London. Intuitively, the fewer worlds an agent considers possible, the less his uncertainty, and the more he knows. If the agent acquires additional information—such as hearing from a reliable source that it is currently sunny in London—then he would no longer consider possible any of the worlds in which it is raining in London.

In a situation such as a poker game, these possible worlds have a concrete interpretation: they are simply all the possible ways the cards could have been distributed among the players. Initially, a player may consider possible all deals consistent with the cards in her hand. Players may acquire additional information in the course of the play of the game that allows them to eliminate some of the worlds they consider possible. Even if Alice does not know originally that Bob holds the ace of spades, at some point Alice might come to know it, if the additional information she obtains allows her to eliminate all the worlds (distributions of cards among players) where Bob does not hold the ace of spades.

Another example is provided by the muddy children puzzle we discussed in the previous chapter. Suppose Alice sees that Bob and Charlie have muddy foreheads and that all the other children do not have muddy foreheads. This allows her to eliminate all but two worlds: one in which she, Bob, and Charlie have muddy foreheads, and no other child does, and one in which Bob and Charlie are the only children with muddy foreheads. In all (i.e., both) of the worlds that Alice considers possible, Bob and Charlie have muddy foreheads and all the children except Bob, Charlie, and her have clean foreheads. Alice's only uncertainty is regarding her own forehead; this uncertainty is reflected in the set of worlds she considers possible. As we shall see in Section 2.3, on hearing the children's replies to the father's first two questions, Alice will be able to eliminate one of these two possible worlds and will know whether or not her own forehead is muddy.

To make these ideas precise, we first need a language that allows us to express notions of knowledge in a straightforward way. As we have already seen, English is not a particularly good language in which to carry out complicated reasoning about knowledge. Instead we use the language of *modal logic*.

Suppose we have a group consisting of n agents, creatively named $1, \ldots, n$. For simplicity, we assume these agents wish to reason about a world that can be described in terms of a nonempty set Φ of *primitive propositions*, typically labeled $p, p', q, q', \ldots$ These primitive propositions stand for basic facts about the world such as "it is sunny in San Francisco" or "Alice has mud on her forehead." To express a statement like "Bob *knows* that it is sunny in San Francisco," we augment the language by *modal* operators $K_1, \ldots, K_n$ (one for each agent). A statement like $K_1\varphi$ is then read "agent 1 knows φ."

Technically, a *language* is just a set of formulas. We can now describe the set of formulas of interest to us. We start with the primitive propositions in Φ and form more complicated formulas by closing off under negation, conjunction, and the modal operators $K_1, \ldots, K_n$. Thus, if φ and ψ are formulas, then so are $\neg\varphi$, $(\varphi \wedge \psi)$, and $K_i\varphi$, for $i = 1, \ldots, n$. For the sake of readability, we omit the parentheses in formulas such as $(\varphi \wedge \psi)$ whenever it does not lead to confusion. We also use standard abbreviations from propositional logic, such as $\varphi \vee \psi$ for $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \Rightarrow \psi$ for $\neg\varphi \vee \psi$, and $\varphi \Leftrightarrow \psi$ for $(\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$. We take *true* to be an abbreviation for some fixed propositional tautology such as $p \vee \neg p$, and take *false* to be an abbreviation for $\neg true$.

We can express quite complicated statements in a straightforward way using this language. For example, the formula

$$K_1 K_2 p \wedge \neg K_2 K_1 K_2 p$$

says that agent 1 knows that agent 2 knows p, but agent 2 does not know that agent 1 knows that agent 2 knows p.

We view possibility as the dual of knowledge. Thus, agent 1 considers φ possible exactly if he does not know $\neg\varphi$. This situation can be described by the formula $\neg K_1 \neg\varphi$. A statement like "Dean doesn't know whether φ" says that Dean considers both φ and $\neg\varphi$ possible. Let's reconsider the sentence from the previous chapter: "Dean doesn't know whether Nixon knows that Dean knows that Nixon knows that McCord burgled O'Brien's office at Watergate." If we take Dean to be agent 1, Nixon to be agent 2, and p to be the statement "McCord burgled O'Brien's office at Watergate," then this sentence can be captured as

$$\neg K_1 \neg(K_2 K_1 K_2 p) \wedge \neg K_1 \neg(\neg K_2 K_1 K_2 p).$$

Now that we have described the *syntax* of our language (that is, the set of well-formed formulas), we need *semantics*, that is, a formal model that we can use to determine whether a given formula is true or false. One approach to defining semantics is, as we suggested, in terms of possible worlds, which we formalize in terms of *(Kripke) structures*. (In later chapters we consider other approaches to giving semantics to formulas.) A Kripke structure M for n agents over Φ is a tuple $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where S is a set of *states* or *possible worlds*, π is an *interpretation* that associates with each state in S a truth assignment to the primitive propositions of Φ (i.e., $\pi(s) : \Phi \to \{\text{true}, \text{false}\}$ for each state $s \in S$), and $\mathcal{K}_i$ is a binary relation on S, that is, a set of pairs of elements of S. We tend to use "state" and "world" interchangeably throughout the book.

The truth assignment $\pi(s)$ tells us whether p is true or false in state s. Thus, if p denotes the fact "it is raining in San Francisco," then $\pi(s)(p) = \textbf{true}$ captures the situation in which it is raining in San Francisco in state s of structure M. The binary relation $\mathcal{K}_i$ is intended to capture the possibility relation according to agent i: $(s, t) \in \mathcal{K}_i$ if agent i considers world t possible, given his information in world s. We think of $\mathcal{K}_i$ as a *possibility* relation, since it defines what worlds agent i considers possible in any given world. Throughout most of the book (in particular, in this chapter), we further require that $\mathcal{K}_i$ be an *equivalence relation* on S. An equivalence relation $\mathcal{K}$ on S is a binary relation that is (a) *reflexive*, which means that for all $s \in S$, we have $(s, s) \in \mathcal{K}$, (b) *symmetric*, which means that for all $s, t \in S$, we have $(s, t) \in \mathcal{K}$ if and only if $(t, s) \in \mathcal{K}$, and (c) *transitive*, which means that for all $s, t, u \in S$, we have that if $(s, t) \in \mathcal{K}$ and $(t, u) \in \mathcal{K}$, then $(s, u) \in \mathcal{K}$. We take $\mathcal{K}_i$ to be an equivalence relation since we want to capture the intuition that agent i considers t possible in world s if in both s and t agent i has the same information about the world, that is, the two worlds are indistinguishable to the agent. Making $\mathcal{K}_i$ an equivalence relation seems natural, and it turns out to be the appropriate choice for many applications. For example, as we shall see in the next section, it is appropriate in analyzing the muddy children puzzle, while in Chapters 4 and 6 we show that it is appropriate for many multi-agent systems applications. We could equally well, however, consider possibility relations with other properties (for example, reflexive and transitive, but not symmetric), as we in fact do in Chapter 3.

We now define what it means for a formula to be true at a given world in a structure. Note that truth depends on the world as well as the structure. It is quite possible that a formula is true in one world and false in another. For example, in one world agent 1 may know it is sunny in San Francisco, while in another he may not. To capture this, we define the notion $(M, s) \models \varphi$, which can be read as "φ *is true at* (M, s)" or "φ *holds at* (M, s)" or "(M, s) *satisfies* φ." We define the $\models$ relation by induction on the structure of φ. That is, we start with the simplest formulas—primitive propositions—and work our way up to more complicated formulas φ, assuming that $\models$ has been defined for all the subformulas of φ.

The π component of the structure gives us the information we need to deal with the base case, where φ is a primitive proposition:

$(M, s) \models p$ (for a primitive proposition $p \in \Phi$) iff $\pi(s)(p) = \textbf{true}$.

For conjunctions and negations, we follow the standard treatment from propositional logic; a conjunction $\psi \wedge \psi'$ is true exactly if both of the conjuncts ψ and ψ' are true, while a negated formula $\neg\psi$ is true exactly if ψ is not true:

$(M, s) \models \psi \wedge \psi'$ iff $(M, s) \models \psi$ and $(M, s) \models \psi'$

$(M, s) \models \neg\psi$ iff $(M, s) \not\models \psi$.

Note that the clause for negation guarantees that the logic is two-valued. For every formula ψ, we have either $(M, s) \models \psi$ or $(M, s) \models \neg\psi$, but not both.

Finally, we have to deal with formulas of the form $K_i\psi$. Here we try to capture the intuition that agent i knows ψ in world s of structure M exactly if ψ is true at all worlds that i considers possible in s. Formally, we have

$(M, s) \models K_i\psi$ iff $(M, t) \models \psi$ for all t such that $(s, t) \in \mathcal{K}_i$.

These definitions are perhaps best illustrated by a simple example. One of the advantages of a Kripke structure is that it can be viewed as a labeled graph, that is, a set of labeled nodes connected by directed, labeled edges. The nodes are the states of S; the label of state $s \in S$ describes which primitive propositions are true and false at s. We label edges by sets of agents; the label on the edge from s to t includes i if $(s, t) \in \mathcal{K}_i$. For example, suppose $\Phi = \{p\}$ and $n = 2$, so that our language has one primitive proposition p and there are two agents. Further suppose that $M = (S, \pi, \mathcal{K}_1, \mathcal{K}_2)$, where $S = \{s, t, u\}$, p is true at states s and u but false at t (so that $\pi(s)(p) = \pi(u)(p) = $ **true** and $\pi(t)(p) = $ **false**), agent 1 cannot distinguish s from t (so that $\mathcal{K}_1 = \{(s, s), (s, t), (t, s), (t, t), (u, u)\}$), and agent 2 cannot distinguish s from u (so that $\mathcal{K}_2 = \{(s, s), (s, u), (t, t), (u, s), (u, u)\}$). This situation can be captured by the graph in Figure 2.1. Note how the graph captures our assumptions about the $\mathcal{K}_i$ relations. In particular, we have a self-loop at each edge labeled by both 1 and 2 because the relations $\mathcal{K}_1$ and $\mathcal{K}_2$ are reflexive, and the edges have an arrow in each direction because $\mathcal{K}_1$ and $\mathcal{K}_2$ are symmetric.
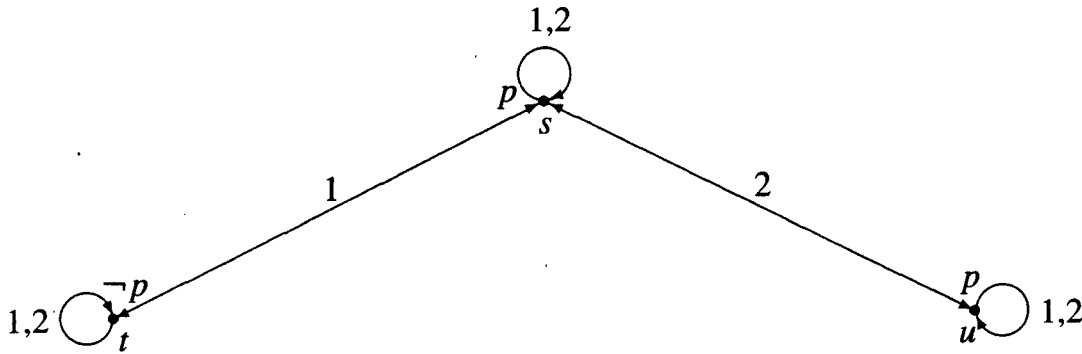

Figure 2.1   A simple Kripke structure

If we view p as standing for "it is sunny in San Francisco," then in state s it is sunny in San Francisco but agent 1 does not know it, since in state s he considers both s and t possible. (We remark that we used the phrase "agent 1 cannot distinguish s from t." Of course, agent 1 realizes perfectly well that s and t are different worlds. After all, it is raining in San Francisco in s, but not in t. What we really intend here is perhaps more accurately described by something like "agent 1's information is insufficient to enable him to distinguish whether the actual world is s or t." We continue to use the word "indistinguishable" in the somewhat looser sense throughout the book.) On the other hand, agent 2 does know in state s that it is sunny, since in both worlds that agent 2 considers possible at s (namely, s and u), the formula p is true. In state t, agent 2 also knows the true situation, namely, that it is not sunny. It follows that in state s agent 1 knows that agent 2 knows whether or not it is sunny in San Francisco: in both worlds agent 1 considers possible in state s, namely, s and t, agent 2 knows what the weather in San Francisco is. Thus, although agent 1 does not know the true situation at s, he does know that agent 2 knows the true situation. By way of contrast, although in state s agent 2 knows that it is sunny in San Francisco, she does not know that agent 1 does not know this fact. (In one world that agent 2 considers possible, namely u, agent 1 does know that it is sunny, while in another world agent 2 considers possible, namely s, agent 1 does not know this fact.) All of this relatively complicated English discussion can be summarized in one mathematical statement:

$$(M, s) \models p \land \neg K_1 p \land K_2 p \land K_1(K_2 p \lor K_2 \neg p) \land \neg K_2 \neg K_1 p.$$

Note that in both s and u, the primitive proposition p (the only primitive proposition in our language) gets the same truth value. One might think, therefore, that s and u are the same, and that perhaps one of them can be eliminated. This is not true! A state is not completely characterized by the truth values that the primitive propositions get there. The possibility relation is also crucial. For example, in world s, agent 1 considers t possible, while in u he does not. As a consequence, agent 1 does not know p in s, while in u he does.

We now consider a slightly more complicated example, which might provide a little more motivation for making the $\mathcal{K}_i$'s equivalence relations. Suppose we have a deck consisting of three cards labeled A, B, and C. Agents 1 and 2 each get one of these cards; the third card is left face down. A possible world is characterized by describing the cards held by each agent. For example, in the world (A, B), agent 1 holds card A and agent 2 holds card B (while card C is face down). There are clearly six possible worlds: (A, B), (A, C), (B, A), (B, C), (C, A), and (C, B). Moreover, it is clear that in a world such as (A, B), agent 1 thinks two worlds are possible:

(A, B) itself and (A, C). Agent 1 knows that he has card A, but he considers it possible that agent 2 could hold either card B or card C. Similarly, in world (A, B), agent 2 also considers two worlds: (A, B) and (C, B). In general, in a world (x, y), agent 1 considers (x, y) and (x, z) possible, and agent 2 considers (x, y) and (z, y) possible, where z is different from both x and y.

From this description, we can easily construct the $\mathcal{K}_1$ and $\mathcal{K}_2$ relations. It is easy to check that they are indeed equivalence relations, as required by the definitions. This is because an agent's possibility relation is determined by the information she has, namely, the card she is holding. This is an important general phenomenon: in any situation where an agent's possibility relation is determined by her information (and, as we shall see, there are many such situations), the possibility relations are equivalence relations.

The structure in this example with the three cards is described in Figure 2.2, where, since the relations are equivalence relations, we omit the self-loops and the arrows on edges for simplicity. (As we have observed, if there is an edge from state s to state t, there is bound to be an edge from t to s as well by symmetry.)
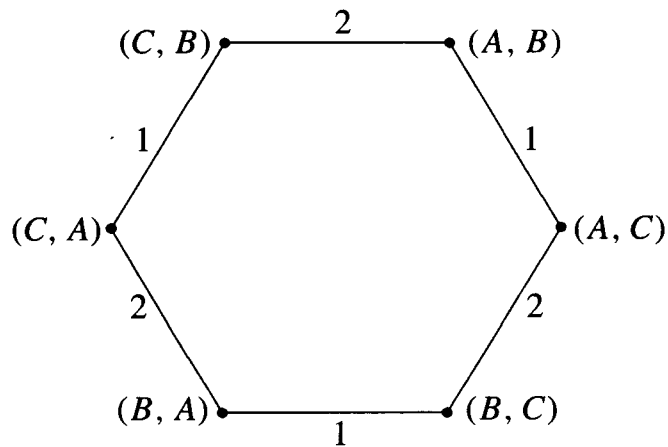

Figure 2.2 The Kripke structure describing a simple card game

This example points out the need for having worlds that an agent does not consider possible included in the structure. For example, in the world (A, B), agent 1 knows that the world (B, C) cannot be the case. (After all, agent 1 knows perfectly well that his own card is A.) Nevertheless, because agent 1 considers it possible that agent 2 considers it possible that (B, C) is the case, we must include (B, C) in the structure. This is captured in the structure by the fact that there is no edge from (A, B)

to (B, C) labeled 1, but there is an edge labeled 1 to (A, C), from which there is an edge labeled 2 to (B, C).

We still have not discussed the language to be used in this example. Since we are interested in reasoning about the cards held by agents 1 and 2, it seems reasonable to have primitive propositions of the form $1A$, $2A$, $2B$, and so on, which are to be interpreted as "agent 1 holds card A," "agent 2 holds card A," "agent 2 holds card B," and so on. Given this interpretation, we define π in the obvious way, and let M_c be the Kripke structure describing this card game. Then, for example, we have $(M_c, (A, B)) \models 1A \wedge 2B$. We leave it to the reader to check that we also have $(M_c, (A, B)) \models K_1(2B \vee 2C)$, which expresses the fact that if agent 1 holds A, then he knows that agent 2 holds either B or C. Similarly, we have $(M_c, (A, B)) \models K_1 \neg K_2(1A)$: agent 1 knows that agent 2 does not know that agent 1 holds A.

This example shows that our semantics does capture some of the intuitions we naturally associate with the word "knowledge." Nevertheless, this is far from a complete justification for our definitions, in particular, for our reading of the formula $K_i \varphi$ as "agent i knows φ." The question arises as to what would constitute a reasonable justification. We ultimately offer two justifications, which we hope the reader will find somewhat satisfactory. The first is by further examples, showing that our definitions correspond to reasonable usages of the word "know." One such example is given in Section 2.3, where we analyze the muddy children puzzle and show that the formula $K_i \varphi$ does capture our intuition regarding what child i knows. The second justification can be found in Section 2.4, where we consider some of the properties of this notion of knowledge and show that they are consistent with the properties that the knowledge of a perfect reasoner with perfect introspection might have. Of course, this does not imply that other reasonable notions of knowledge do not exist. Some of these are considered in later chapters.

We have also restricted attention here to *propositional* modal logic. We do not have first-order quantification, so that we cannot easily say, for example, that Alice knows the governors of all states. Such a statement would require universal and existential quantification. Roughly speaking, we could express it as $\forall x (State(x) \Rightarrow \exists y (K_{Alice}Governor(x, y))$: for all states x, there exists y such that Alice knows that the governor of x is y. We use only propositional modal logic throughout most of this book because it is sufficiently powerful to capture most of the situations we shall be interested in, while allowing us to avoid some of the complexities that arise in the first-order case. We briefly consider the first-order case in Section 3.7.

2.2   Adding Common Knowledge and Distributed Knowledge

The language introduced in the previous section does not allow us to express the notions of common knowledge and distributed knowledge that we discussed in Chapter 1. To express these notions, we augment the language with the modal operators E_G ("everyone in the group G knows"), C_G ("it is common knowledge among the agents in G"), and D_G ("it is distributed knowledge among the agents in G") for every nonempty subset G of $\{1, \ldots, n\}$, so that if φ is a formula, then so are $E_G\varphi$, $C_G\varphi$, and $D_G\varphi$. We often omit the subscript G when G is the set of all agents. In this augmented language we can make statements like $K_3 \neg C_{\{1,2\}} p$ ("agent 3 knows that p is not common knowledge among agents 1 and 2") and $Dq \wedge \neg Cq$ ("q is distributed knowledge, but it is not common knowledge").

We can easily extend the definition of truth to handle common knowledge and distributed knowledge in a structure M. Since $E_G\varphi$ is true exactly if everyone in the group G knows φ, we have

$$(M, s) \models E_G\varphi \text{ iff } (M, s) \models K_i\varphi \text{ for all } i \in G.$$

The formula $C_G\varphi$ is true if everyone in G knows φ, everyone in G knows that everyone in G knows φ, etc. Let $E_G^0\varphi$ be an abbreviation for φ, and let $E_G^{k+1}\varphi$ be an abbreviation for $E_G E_G^k\varphi$. In particular, $E_G^1\varphi$ is an abbreviation for $E_G\varphi$. Then we have

$$(M, s) \models C_G\varphi \text{ iff } (M, s) \models E_G^k\varphi \text{ for } k = 1, 2, \ldots$$

Our definition of common knowledge has an interesting graph-theoretical interpretation, which turns out to be useful in many of our applications. Define a state t to be *G-reachable from state s in k steps* ($k \geq 1$) if there exist states $s_0, s_1, \ldots, s_k$ such that $s_0 = s$, $s_k = t$ and for all j with $0 \leq j \leq k - 1$, there exists $i \in G$ such that $(s_j, s_{j+1}) \in \mathcal{K}_i$. We say t is *G-reachable from s* if t is G-reachable from s in k steps for some $k \geq 1$. Thus, t is G-reachable from s exactly if there is a path in the graph from s to t whose edges are labeled by members of G. In the particular case where G is the set of all agents, we say simply that t is *reachable from s*. Thus, t is reachable from s exactly if s and t are in the same connected component of the graph.

Lemma 2.2.1

(a) $(M, s) \models E_G^k\varphi$ if and only if $(M, t) \models \varphi$ for all t that are G-reachable from s in k steps.

(b) $(M, s) \models C_G\varphi$ if and only if $(M, t) \models \varphi$ for all t that are G-reachable from s.

Proof   Part (a) follows from a straightforward induction on k, while part (b) is immediate from part (a). Notice that this result holds even if the $\mathcal{K}_i$'s are arbitrary binary relations; we do not need to assume that they are equivalence relations. ∎

A group G has distributed knowledge of φ if the "combined" knowledge of the members of G implies φ. How can we capture the idea of combining knowledge in our framework? In the Kripke structure in Figure 2.1, in state s agent 1 considers both s and t possible but does not consider u possible, while agent 2 considers s and u possible, but not t. Someone who could combine the knowledge of agents 1 and 2 would know that only s was possible: agent 1 has enough knowledge to eliminate u, and agent 2 has enough knowledge to eliminate t. In general, we combine the knowledge of the agents in group G by eliminating all worlds that some agent in G considers impossible. Technically, this is accomplished by *intersecting* the sets of worlds that each of the agents in the group considers possible. Thus we define

$$(M, s) \models D_G\varphi \text{ iff } (M, t) \models \varphi \text{ for all } t \text{ such that } (s, t) \in \cap_{i \in G}\mathcal{K}_i.$$

Returning to our card game example, let $G = \{1, 2\}$; thus, G is the group consisting of the two players in the game. Then it is easy to check (using Lemma 2.2.1) that $(M_c, (A, B)) \models C_G(1A \vee 1B \vee 1C)$: it is common knowledge that agent 1 holds one of the cards A, B, and C. Perhaps more interesting is $(M_c, (A, B)) \models C_G(1B \Rightarrow (2A \vee 2C))$: it is common knowledge that if agent 1 holds card B, then agent 2 holds either card A or card C. More generally, it can be shown that any fact about the game that can be expressed in terms of the propositions in our language is common knowledge.

What about distributed knowledge? We leave it to the reader to check that, for example, we have $(M_c, (A, B)) \models D_G(1A \wedge 2B)$. If the agents could pool their knowledge together, they would know that in world (A, B), agent 1 holds card A and agent 2 holds card B.

Again, this example does not provide complete justification for our definitions. But it should at least convince the reader that they are plausible. We examine the properties of common knowledge and distributed knowledge in more detail in Section 2.4.

2.3   The Muddy Children Revisited

In our analysis we shall assume that it is common knowledge that the father is truthful, that all the children can and do hear the father, that all the children can and do see

which of the other children besides themselves have muddy foreheads, that none of the children can see his or her own forehead, and that all the children are truthful and (extremely) intelligent.

First consider the situation before the father speaks. Suppose there are n children altogether. As before, we number them $1, \ldots, n$. Some of the children have muddy foreheads, and the rest do not. We can describe a possible situation by an n-tuple of 0's and 1's of the form $(x_1, \ldots, x_n)$, where $x_i = 1$ if child i has a muddy forehead, and $x_i = 0$ otherwise. Thus, if $n = 3$, then a tuple of the form $(1, 0, 1)$ would say that precisely child 1 and child 3 have muddy foreheads. Suppose the actual situation is described by this tuple. What situations does child 1 consider possible before the father speaks? Since child 1 can see the foreheads of all the children besides herself, her only doubt is about whether she has mud on her own forehead. Thus child 1 considers two situations possible, namely, $(1, 0, 1)$ (the actual situation) and $(0, 0, 1)$. Similarly, child 2 considers two situations possible: $(1, 0, 1)$ and $(1, 1, 1)$. Note that in general, child i has the same information in two possible worlds exactly if they agree in all components except possibly the ith component.

We can capture the general situation by a Kripke structure M consisting of 2^n states, one for each of the possible n-tuples. We must first decide what propositions we should include in our language. Since we want to reason about whether or not a given child's forehead is muddy, we take $\Phi = \{p_1, \ldots, p_n, p\}$, where, intuitively, p_i stands for "child i has a muddy forehead," and p stands for "at least one child has a muddy forehead." Thus, we define π so that $(M, (x_1, \ldots, x_n)) \models p_i$ if and only if $x_i = 1$, and $(M, (x_1, \ldots, x_n)) \models p$ if and only if $x_j = 1$ for some j. Of course, p is equivalent to $p_1 \vee \ldots \vee p_n$, so its truth value can be determined from the truth value of the other primitive propositions. There is nothing to prevent us from choosing a language where the primitive propositions are not independent. Since it is convenient to add a primitive proposition (namely p) describing the father's statement, we do so. Finally, we must define the $\mathcal{K}_i$ relations. Since child i considers a world possible if it agrees in all components except possibly the ith component, we take $(s, t) \in \mathcal{K}_i$ exactly if s and t agree in all components except possibly the ith component. Notice that this definition makes $\mathcal{K}_i$ an equivalence relation. This completes the description of M.

Although this Kripke structure may seem quite complicated, it actually has an elegant graphical representation. Suppose we ignore self-loops and the labeling on the edges for the moment. Then we have a structure with 2^n nodes, each described by an n-tuple of 0's and 1's, such that two nodes are joined by an edge exactly if they differ in one component. The reader with a good imagination will see that this
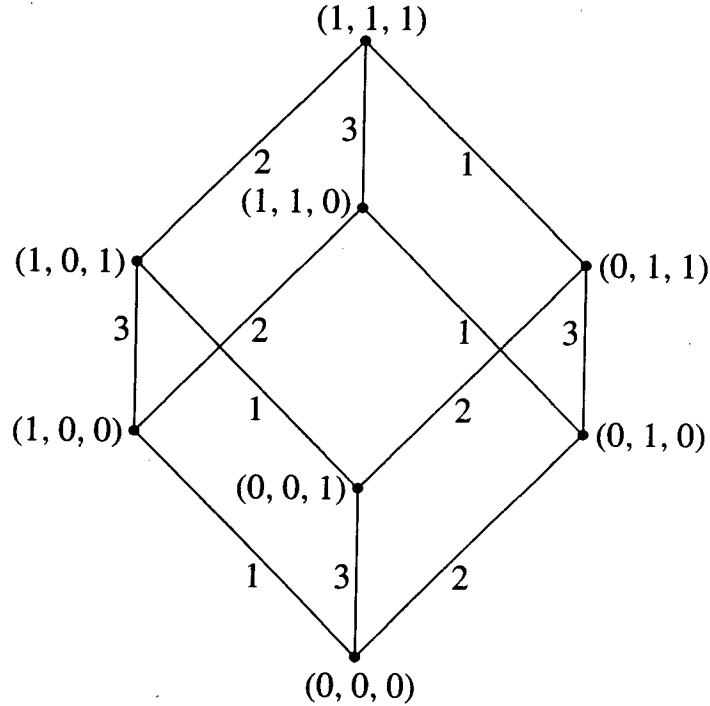
Figure 2.3   The Kripke structure for the muddy children puzzle with $n = 3$

defines an n-dimensional cube. The case $n = 3$ is illustrated in Figure 2.3 (where again we omit self-loops and the arrows on edges).

Intuitively, each child knows which of the other children have muddy foreheads. This intuition is borne out in our formal definition of knowledge. For example, it is easy to see that when the actual situation is $(1, 0, 1)$, we have $(M, (1, 0, 1)) \models K_1 \neg p_2$, since when the actual situation is $(1, 0, 1)$, child 2 does not have a muddy forehead in both worlds that child 1 considers possible. Similarly, we have $(M, (1, 0, 1)) \models K_1 p_3$: child 1 knows that child 3's forehead is muddy. However, $(M, (1, 0, 1)) \models \neg K_1 p_1$. Child 1 does not know that her own forehead is muddy, since in the other world she considers possible—$(0,0,1)$—her forehead is not muddy. In fact, it is common knowledge that every child knows whether every other child's forehead is muddy or not. Thus, for example, a formula like $p_2 \Rightarrow K_1 p_2$, which says that if child 2's forehead is muddy then child 1 knows it, is common knowledge. We leave it to the reader to check that $C(p_2 \Rightarrow K_1 p_2)$ is true at every state, as is $C(\neg p_2 \Rightarrow K_1 \neg p_2)$.

In the world $(1,0,1)$, in which there are two muddy children, every child knows that at least one child has a muddy forehead even before the father speaks. And sure enough, we have $(M, (1, 0, 1)) \models Ep$. It follows, however, from Lemma 2.2.1 that

$(M, (1, 0, 1)) \models \neg E^2 p$, since p is not true at the world $(0, 0, 0)$ that is reachable in two steps from $(1, 0, 1)$. The reader can easily check that in the general case, if we have n children of whom k have muddy foreheads (so that the situation is described by an n-tuple exactly k of whose components are 1's), then $E^{k-1} p$ is true, but $E^k p$ is not, since each world (tuple) reachable in $k - 1$ steps has at least one 1 (and so there is at least one child with a muddy forehead), but the tuple $(0, \ldots, 0)$ is reachable in k steps.

Before we go on, the reader should note that there are a number of assumptions implicit in our representation. The fact that we have chosen to represent a world as an n-tuple in this way is legitimate if we can assume that all the information necessary for our reasoning already exists in such tuples. If there were some doubt as to whether child 1 was able to see, then we would have to include this information in the state description as well. Note also that the assumption that it is common knowledge that all the children can see is what justifies the choice of edges. For example, if $n = 3$ and if it were common knowledge that child 1 is blind, then, for example, in the situation $(1, 1, 1)$, child 1 would also consider $(1, 0, 0)$ possible. He would not know that child 2's forehead is muddy (see Exercises 2.1 and 2.2).

In general, when we choose to model a given situation, we have to put into the model everything that is relevant. One obvious reason that a fact may be "irrelevant" is because it does not pertain to the situation we are analyzing. Thus, for example, whether child 1 is a boy or a girl is not part of the description of the possible world. Another cause of irrelevance is that a fact may be common knowledge. If it is common knowledge that all the children can see, then there is no point in adding this information to the description of a possible world. It is true at all the possible worlds in the picture, so we do not gain anything extra by mentioning it. Thus, common knowledge can help to simplify our description of a situation.

We remark that throughout the preceding discussion we have used the term "common knowledge" in two slightly different, although related, senses. The first is the technical sense, where a formula φ in our language is common knowledge at a state s if it is true at all states reachable from s. The second is a somewhat more informal sense, where we say that a fact (not necessarily expressible in our language) is common knowledge if it is true at all the situations (states) in the structure. When we say it is common knowledge that at least one child has mud on his or her forehead, then we are using common knowledge in the first sense, since this corresponds to the formula Cp. When we say that it is common knowledge that no child is blind, we are using it in the second sense, since we do not have a formula q in the language that says that no child is blind. There is an obvious relationship between the two senses of the term. For example, if we enrich our language so that it does have a formula q

saying "no child is blind," then Cq actually would hold at every state in the Kripke structure. Throughout this book, we continue to speak of common knowledge in both senses of the term, and we hope that the reader can disambiguate if necessary.

Returning to our analysis of the puzzle, consider what happens after the father speaks. The father says p, which, as we have just observed, is already known to all the children if there are two or more children with muddy foreheads. Nevertheless, the state of knowledge changes, even if all the children already know p. If $n = 3$, in the world $(1, 0, 1)$ child 1 considers the situation $(0, 0, 1)$ possible. In that world, child 3 considers $(0, 0, 0)$ possible. Thus, in the world $(1, 0, 1)$, before the father speaks, although everyone knows that at least one child has a muddy forehead, child 1 thinks it possible that child 3 thinks it possible that none of the children has a muddy forehead. After the father speaks, it becomes *common knowledge* that at least one child has a muddy forehead. (This, of course, depends on our assumption that it is common knowledge that all the children can and do hear the father.) We can represent the change in the group's state of knowledge graphically (in the general case) by simply removing the point $(0, 0, \ldots, 0)$ from the cube, getting a "truncated" cube. (More accurately, what happens is that the node $(0, 0, \ldots, 0)$ remains, but all the edges between $(0, 0, \ldots, 0)$ and nodes with exactly one 1 disappear, since it is common knowledge that even if only one child has a muddy forehead, after the father speaks that child will not consider it possible that no one has a muddy forehead.) The situation is illustrated in Figure 2.4.
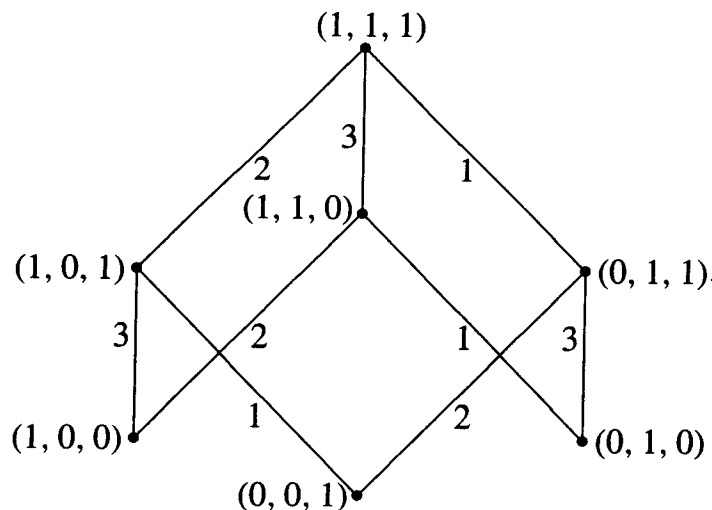

Figure 2.4   The Kripke structure after the father speaks

We next show that each time the children respond to the father's question with a "No", the group's state of knowledge changes and the cube is further truncated. Consider what happens after the children respond "No" to the father's first question. We claim that now all the nodes with exactly one 1 can be eliminated. (More accurately, the edges to these nodes from nodes with exactly two 1's all disappear from the graph.) Nodes with one or fewer 1's are no longer reachable from nodes with two or more 1's. The reasoning here parallels the reasoning in the "proof" given in the story. If the actual situation were described by, say, the tuple $(1, 0, \ldots, 0)$, then child 1 would initially consider two situations possible: $(1, 0, \ldots, 0)$ and $(0, 0, \ldots, 0)$. Once the father speaks it is common knowledge that $(0, 0, \ldots, 0)$ is not possible, so he would then know that the situation is described by $(1, 0, \ldots, 0)$, and thus would know that his own forehead is muddy. Once everyone answers "No" to the father's first question, it is common knowledge that the situation cannot be $(1, 0, \ldots, 0)$. (Note that here we must use the assumption that it is common knowledge that everyone is intelligent and truthful, and so can do the reasoning required to show $(1, 0, \ldots, 0)$ is not possible.) Similar reasoning allows us to eliminate every situation with exactly one 1. Thus, after all the children have answered "No" to the father's first question, it is common knowledge that at least *two* children have muddy foreheads.

Further arguments in the same spirit can be used to show that after the children answer "No" k times, we can eliminate all the nodes with at most k 1's (or, more accurately, disconnect these nodes from the rest of the graph). We thus have a sequence of Kripke structures, describing the children's knowledge at every step in the process. Essentially, what is going on is that if, in some node s, it becomes common knowledge that a node t is impossible, then for every node u reachable from s, the edge from u to t (if there is one) is eliminated. (This situation is even easier to describe once we add time to the picture. We return to this point in Chapter 7; see in particular Section 7.2.)

After k rounds of questioning, it is common knowledge that at least $k+1$ children have mud on their foreheads. If the true situation is described by a tuple with exactly $k+1$ entries of 1, then before the father asks the question for the $(k+1)^{\text{st}}$ time, those children with muddy foreheads will know the exact situation, and in particular will know their foreheads are muddy, and consequently will answer "Yes." Note that they could not answer "Yes" any earlier, since up to this point each child with a muddy forehead considers it possible that he or she does not have a muddy forehead.

There is actually a subtle point that should be brought out here. Roughly speaking, according to the way we are modeling "knowledge" in this context, a child "knows" a fact if the fact follows from his or her current information. But we could certainly imagine that if one of the children were not particularly bright, then he might not

be able to figure out that he "knew" that his forehead was muddy, even though in principle he had enough information to do so. To answer "Yes" to the father's question, it really is not enough for it to follow from the child's information whether the child has a muddy forehead. The child must actually be aware of the consequences of his information—that is, in some sense, the child must be able to compute that he has this knowledge—in order to act on it. Our definition implicitly assumes that all reasoners are *logically omniscient*, i.e., they are smart enough to compute all the consequences of the information that they have.

Now consider the situation in which the father does not initially say p. We claim that in this case the children's state of knowledge never changes, no matter how many times the father asks questions. It can always be described by the n-dimensional cube. We have already argued that before the father speaks the situation is described by the n-dimensional cube. When the father asks for the first time "Do any of you know whether you have mud on your own forehead?," clearly all the children say "No," no matter what the actual situation is, since in every situation each child considers possible a situation in which he or she does not have a muddy forehead. Since it is common knowledge before the father asks his question that the answer will be "No", no information is gained from this answer, so the situation still can be represented by the n-dimensional cube. Now a straightforward induction on m shows that it is common knowledge that the father's m^{th} question is also answered "No" (since at the point when the father asks this question, no matter what the situation is, each child will consider possible another situation in which he does not have a muddy forehead), and the state of knowledge after the father asks the mth question is still described by the cube.

This concludes our analysis of the muddy children puzzle.

2.4   The Properties of Knowledge

In the first part of this chapter we described a language with modal operators such as K_i and defined a notion of truth that, in particular, determines whether a formula such as $K_i\varphi$ is true at a particular world. We suggested that $K_i\varphi$ should be read as "agent i knows φ." But is this a reasonable way of reading this formula? Does our semantics—that is, Kripke structures together with the definition of truth that we gave—really capture the properties of knowledge in a reasonable way? How can we even answer this question?

We can attempt to answer the question by examining what the properties of knowledge are under our interpretation. One way of characterizing the properties

of our interpretation of knowledge is by characterizing the formulas that are always true. More formally, given a structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, we say that φ is *valid in M*, and write $M \models \varphi$, if $(M, s) \models \varphi$ for every state s in S, and we say that φ is *satisfiable in M* if $(M, s) \models \varphi$ for some state s in S. We say that φ is *valid*, and write $\models \varphi$, if φ is valid in all structures, and that φ is *satisfiable* if it is satisfiable in some structure. It is easy to check that a formula φ is valid (resp. valid in M) if and only if $\neg\varphi$ is not satisfiable (resp. not satisfiable in M).

We now list a number of valid properties of our definition of knowledge and provide a formal proof of their validity. We then discuss how reasonable these properties are. As before, we assume throughout this section that the possibility relations $\mathcal{K}_i$ are equivalence relations.

One important property of our definition of knowledge is that each agent knows all the logical consequences of his knowledge. If an agent knows φ and knows that φ implies ψ, then both φ and $\varphi \Rightarrow \psi$ are true at all worlds he considers possible. Thus ψ must be true at all worlds that the agent considers possible, so he must also know ψ. It follows that

$$\models (K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi.$$

This axiom is called the *Distribution Axiom* since it allows us to distribute the K_i operator over implication. It seems to suggest that our agents are quite powerful reasoners.

Further evidence that our definition of knowledge assumes rather powerful agents comes from the fact that agents know all the formulas that are valid in a given structure. If φ is true at all the possible worlds of structure M, then φ must be true at all the worlds that an agent considers possible in any given world in M, so it must be the case that $K_i\varphi$ is true at all possible worlds of M. More formally, we have the following *Knowledge Generalization Rule*:

For all structures M, if $M \models \varphi$, then $M \models K_i\varphi$.

Note that from this we can deduce that if φ is valid, then so is $K_i\varphi$. This rule is very different from the formula $\varphi \Rightarrow K_i\varphi$, which says that if φ is true, then agent i knows it. An agent does not necessarily know all things that are true. (For example, in the case of the muddy children, it may be true that child 1 has a muddy forehead, but he does not necessarily know this.) However, agents do know all valid formulas. Intuitively, these are the formulas that are *necessarily* true, as opposed to the formulas that just happen to be true at a given world.

Although an agent may not know facts that are true, it is the case that if an agent knows a fact, then it is true. More formally, we have

$$\models K_i\varphi \Rightarrow \varphi.$$

This property, occasionally called the *Knowledge Axiom* or the *Truth Axiom* (for knowledge), has been taken by philosophers to be the major one distinguishing knowledge from *belief.* Although you may have false beliefs, you cannot know something that is false. This property follows because the actual world is always one of the worlds that an agent considers possible. If $K_i\varphi$ holds at a particular world (M, s), then φ is true at all worlds that i considers possible, so in particular it is true at (M, s).

The last two properties we consider say that agents can do introspection regarding their knowledge. They know what they know and what they do not know:

$$\models K_i\varphi \Rightarrow K_iK_i\varphi,$$

$$\models \neg K_i\varphi \Rightarrow K_i\neg K_i\varphi.$$

The first of these properties is typically called the *Positive Introspection Axiom*, and the second is called the *Negative Introspection Axiom*.

The following theorem provides us with formal assurance that all the properties just discussed hold for our definition of knowledge.

Theorem 2.4.1   *For all formulas φ and ψ, all structures M where each possibility relation $\mathcal{K}_i$ is an equivalence relation, and all agents $i = 1, \ldots, n$,*

(a) $M \models (K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi,$

(b) *if $M \models \varphi$, then $M \models K_i\varphi$,*

(c) $M \models K_i\varphi \Rightarrow \varphi,$

(d) $M \models K_i\varphi \Rightarrow K_iK_i\varphi,$

(e) $M \models \neg K_i\varphi \Rightarrow K_i\neg K_i\varphi.$

Proof

(a) If $(M, s) \models K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)$, then for all states t such that $(s, t) \in \mathcal{K}_i$, we have both that $(M, t) \models \varphi$ and $(M, t) \models \varphi \Rightarrow \psi$. By the definition of $\models$, we have that $(M, t) \models \psi$ for all such t, and therefore $(M, s) \models K_i\psi$.

(b) If $M \models \varphi$, then $(M, t) \models \varphi$ for all states t in M. In particular, for any fixed state s in M, it follows that $(M, t) \models \varphi$ for all t such that $(s, t) \in \mathcal{K}_i$. Thus, $(M, s) \models K_i \varphi$ for all states s in M, and hence $M \models K_i \varphi$.

(c) If $(M, s) \models K_i \varphi$, then for all t such that $(s, t) \in \mathcal{K}_i$, we have $(M, t) \models \varphi$. Since $\mathcal{K}_i$ is reflexive, it follows that $(s, s) \in \mathcal{K}_i$, so in particular $(M, s) \models \varphi$.

(d) Suppose $(M, s) \models K_i \varphi$. Consider any t such that $(s, t) \in \mathcal{K}_i$ and any u such that $(t, u) \in \mathcal{K}_i$. Since $\mathcal{K}_i$ is transitive, we have $(s, u) \in \mathcal{K}_i$. Since $(M, s) \models K_i \varphi$, it follows that $(M, u) \models \varphi$. Thus, for all t such that $(s, t) \in \mathcal{K}_i$, we have $(M, t) \models K_i \varphi$. It now follows that $(M, s) \models K_i K_i \varphi$.

(e) Suppose $(M, s) \models \neg K_i \varphi$. Then for some u with $(s, u) \in \mathcal{K}_i$, we must have $(M, u) \models \neg \varphi$. Suppose t is such that $(s, t) \in \mathcal{K}_i$. Since $\mathcal{K}_i$ is symmetric, $(t, s) \in \mathcal{K}_i$, and since $\mathcal{K}_i$ is transitive, we must also have $(t, u) \in \mathcal{K}_i$. Thus it follows that $(M, t) \models \neg K_i \varphi$. Since this is true for all t such that $(s, t) \in \mathcal{K}_i$, we obtain $(M, s) \models K_i \neg K_i \varphi$. $\blacksquare$

The collection of properties that we have considered so far—the Distribution Axiom, the Knowledge Axiom, the Positive and Negative Introspection Axioms, and the Knowledge Generalization Rule—has been studied in some depth in the literature. For historical reasons, these properties are sometimes called the *S5 properties*. (Actually, S5 is an axiom system. We give a more formal definition of it in the next chapter.) How reasonable are these properties? The proof of Theorem 2.4.1 shows that, in a precise sense, the validity of the Knowledge Axiom follows from the fact that $\mathcal{K}_i$ is reflexive, the validity of the Positive Introspection Axiom follows from the fact that $\mathcal{K}_i$ is transitive, and the validity of the Negative Introspection Axiom follows from the fact that $\mathcal{K}_i$ is symmetric and transitive. Although taking $\mathcal{K}_i$ to be an equivalence relation seems reasonable for many applications we have in mind, we can certainly imagine other possibilities. As we show in Chapter 3, by modifying the properties of the $\mathcal{K}_i$ relations, we can get notions of knowledge that have different properties.

Two properties that seem forced on us by the possible-worlds approach itself are the Distribution Axiom and the Knowledge Generalization Rule. No matter how we modify the $\mathcal{K}_i$ relations, these properties hold. (This is proved formally in the next chapter.) These properties may be reasonable if we identify "agent i knows φ" with "φ follows from agent i's information," as we implicitly did when modeling the muddy children puzzle. To the extent that we think of knowledge as something acquired by agents through some reasoning process, these properties suggest that we must think

in terms of agents who can do perfect reasoning. While this may be a reasonable idealization in certain circumstances (and is an assumption that is explicitly made in the description of the muddy children puzzle), it is clearly not so reasonable in many contexts. In Chapters 9 and 10 we discuss how the possible-worlds model can be modified to accommodate imperfect, "non-ideal" reasoners.

The reader might wonder at this point if there are other important properties of our definition of knowledge that we have not yet mentioned. Although, of course, a number of additional properties follow from the basic S5 properties, in a precise sense the S5 properties completely characterize our definition of knowledge, at least as far as the K_i operators are concerned. This point is discussed in detail in Chapter 3.

We now turn our attention to the properties of the operators E_G, C_G, and D_G. Since $E_G\varphi$ is true exactly if every agent in G knows φ, we have

$$\models E_G\varphi \Leftrightarrow \bigwedge_{i \in G} K_i\varphi.$$

Recall that we said common knowledge could be viewed as what "any fool" knows. Not surprisingly, it turns out that common knowledge has all the properties of knowledge; axioms analogous to the Knowledge Axiom, Distribution Axiom, Positive Introspection Axiom, and Negative Introspection Axiom all hold for common knowledge (see Exercise 2.8). In addition, it is easy to see that common knowledge among a group of agents implies common knowledge among any of its subgroups, that is, $C_G\varphi \Rightarrow C_{G'}\varphi$ if $G \supseteq G'$ (again, see Exercise 2.8). It turns out that all these properties follow from two other properties, which in a precise sense capture the essence of common knowledge. We discuss these properties next.

Recall from Chapter 1 that the children in the muddy children puzzle acquire common knowledge of the fact p (that at least one child has a muddy forehead) because the father's announcement puts them in a situation where all the children know both that p is true and that they are in this situation. This observation is generalized in the following *Fixed-Point Axiom*, which says that φ is common knowledge among the group G if and only if all the members of G know that φ is true and is common knowledge:

$$\models C_G\varphi \Leftrightarrow E_G(\varphi \land C_G\varphi).$$

Thus, the Fixed-Point Axiom says that $C_G\varphi$ can be viewed as a *fixed point* of the function $f(x) = E_G(\varphi \land x)$, which maps a formula x to the formula $E_G(\varphi \land x)$. (We shall see a formalization of this intuition in Section 11.5.)

The second property of interest gives us a way of deducing that common knowledge holds in a structure.

For all structures M, if $M \models \varphi \Rightarrow E_G(\psi \land \varphi)$, then $M \models \varphi \Rightarrow C_G\psi$.

This rule is often called the *Induction Rule*. The proof that it holds shows why: the antecedent gives us the essential ingredient for proving, by induction on k, that $\varphi \Rightarrow E^k(\psi \wedge \varphi)$ is valid for all k.

We now prove formally that these properties do indeed hold for the operators E_G and C_G.

Theorem 2.4.2 *For all formulas φ and ψ, all structures M, and all nonempty $G \subseteq \{1, \ldots, n\}$,*

(a) $M \models E_G\varphi \Leftrightarrow \bigwedge_{i \in G} K_i\varphi$,

(b) $M \models C_G\varphi \Leftrightarrow E_G(\varphi \wedge C_G\varphi)$,

(c) *if $M \models \varphi \Rightarrow E_G(\psi \wedge \varphi)$, then $M \models \varphi \Rightarrow C_G\psi$.*

Proof Part (a) follows immediately from the semantics of E_G. To prove the other parts, we use the characterization of common knowledge provided by Lemma 2.2.1, namely, that $(M, s) \models C_G\varphi$ iff $(M, t) \models \varphi$ for all states t that are G-reachable from s. We remark for future reference that the proof we are about to give does not make use of the fact that the $\mathcal{K}_i$'s are equivalence relations; it goes through without change even if the $\mathcal{K}_i$'s are arbitrary binary relations.

For part (b), suppose $(M, s) \models C_G\varphi$. Thus $(M, t) \models \varphi$ for all states t that are G-reachable from s. In particular, if u is G-reachable from s in one step, then $(M, u) \models \varphi$ and $(M, t) \models \varphi$ for all t that are G-reachable from u. Thus $(M, u) \models \varphi \wedge C_G\varphi$ for all u that are G-reachable from s in one step, so $(M, s) \models E_G(\varphi \wedge C_G\varphi)$. For the converse, suppose $(M, s) \models E_G(\varphi \wedge C_G\varphi)$. Suppose that t is G-reachable from s and s' is the first node after s on a path from s to t whose edges are labeled by members of G. Since $(M, s) \models E_G(\varphi \wedge C_G\varphi)$, it follows that $(M, s') \models \varphi \wedge C_G\varphi$. Either $s' = t$ or t is reachable from s'. In the former case, $(M, t) \models \varphi$ since $(M, s') \models \varphi$, while in the latter case, $(M, t) \models \varphi$ by Lemma 2.2.1 and the fact that $(M, s') \models C_G\varphi$. Because $(M, t) \models \varphi$ for all t that are G-reachable from s, it follows that $(M, s) \models C_G\varphi$.

Finally, for part (c), suppose $M \models \varphi \Rightarrow E_G(\psi \wedge \varphi)$ and $(M, s) \models \varphi$. We show by induction on k that for all k we have $(M, t) \models \psi \wedge \varphi$ for all t that are G-reachable from s in k steps. Suppose t is G-reachable from s in one step. Since $M \models \varphi \Rightarrow E_G(\psi \wedge \varphi)$, we have $(M, s) \models E_G(\psi \wedge \varphi)$. Since t is G-reachable from s in one step, by Lemma 2.2.1, we have $(M, t) \models \psi \wedge \varphi$ as desired. If $k = k' + 1$, then there is some t' that is G-reachable from s in k' steps such that t is G-reachable from t' in one step. By the induction hypothesis, we have $(M, t') \models \psi \wedge \varphi$. Now the same argument as in the base case shows that $(M, t) \models \psi \wedge \varphi$. This completes

the inductive proof. Since $(M, t) \models \psi$ for all states t that are G-reachable from s, it follows that $(M, s) \models C_G\varphi$. ∎

Finally, we consider distributed knowledge. We mentioned in Chapter 1 that distributed knowledge can be viewed as what a "wise man" would know. So it should not be surprising that distributed knowledge also satisfies all the properties of knowledge. Distributed knowledge has two other properties that we briefly mention here. Clearly, distributed knowledge of a group of size one is the same as knowledge, so that we have

$$\models D_{\{i\}}\varphi \Leftrightarrow K_i\varphi.$$

The larger the subgroup, the greater the distributed knowledge of that subgroup:

$$\models D_G\varphi \Rightarrow D_{G'}\varphi \text{ if } G \subseteq G'.$$

The proof that all these properties of distributed knowledge are indeed valid is similar in spirit to the proof of Theorem 2.4.1, so we leave it to the reader (Exercise 2.10). We also show in Chapter 3 that these properties of common knowledge and distributed knowledge in a precise sense completely characterize all the relevant properties of these notions.

2.5   An Event-Based Approach

The approach to modeling knowledge presented in Section 2.1 has two components. It uses Kripke structures as a mathematical model for situations involving many agents, and it uses a logical language to make assertions about such situations. This language is based on a set of primitive propositions and is closed under logical operators. Thus, knowledge is expressed syntactically, by modal operators on formulas. We call this the *logic-based approach*. It is the approach that traditionally has been taken in philosophy, mathematical logic, and AI.

In this section, we describe an alternate approach to modeling knowledge, one that is typically used in the work on knowledge in game theory and mathematical economics. We call this the *event-based approach*. It differs from the logic-based approach in two respects. First, rather than use Kripke structures as the underlying mathematical model, the event-based approach uses closely related structures that we call *Aumann structures*. Second, and more important, in the spirit of probability theory, the event-based approach focuses on *events*, which are sets of possible worlds, and dispenses completely with logical formulas. Knowledge here is expressed as an

operator on events. We now review the event-based approach and discuss its close relationship to the logic-based approach.

As in the logic-based approach of Section 2.1, we start out with a universe S of *states*. An *event* is a set $e \subseteq S$ of states. We can talk, for example, about the event of its raining in London, which corresponds to the set of states where it is raining in London. We say that *event e holds at state s* if $s \in e$. Thus, if e_L is the event of its raining in London, then e_L holds at state s precisely if s is one of the states where it is raining in London. The conjunction of two events is given by their intersection. For example, the event of its raining in London and being sunny in San Francisco is the intersection of e_L with the event of its being sunny in San Francisco. Similarly, the negation of an event is given by the complement (with respect to S).

As we have mentioned, Aumann structures are used to provide a formal model for the event-based approach. Aumann structures are like Kripke structures, with two differences: The first is that there is no analogue to the π function, since in the event-based approach, there are no primitive propositions. The second difference is that, rather than use a binary relation $\mathcal{K}_i$ to define what worlds agent i considers possible, in Aumann structures there is a *partition* $\mathcal{P}_i$ of S for each agent i. (A *partition* of a set S is a set $\{S_1, \ldots, S_r\}$ of subsets of S such that the S_j's are disjoint and such that the union of the S_j's is the set S.) If $\mathcal{P}_i = \{S_1, \ldots, S_r\}$, then the sets S_j are called the *cells* of the partition $\mathcal{P}_i$, or the *information sets* of agent i. The intuition is that if S_j is an information set of agent i, and if $s \in S_j$, then the set of states that agent i considers possible (which corresponds to the information of agent i) is precisely S_j.

Formally, an Aumann structure A is a tuple $(S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$, where S is the set of states of the world and $\mathcal{P}_i$ is a partition of S for every agent i. We denote by $\mathcal{P}_i(s)$ the cell of the partition $\mathcal{P}_i$ in which s appears. Since $\mathcal{P}_i$ is a partition, it follows that for every agent i and every pair $s, t \in S$ of states, either $\mathcal{P}_i(s) = \mathcal{P}_i(t)$ or $\mathcal{P}_i(s) \cap \mathcal{P}_i(t) = \emptyset$. Intuitively, when s, t are in the same information set of agent i, then in state s agent i considers the state t possible. As we have already remarked, unlike a Kripke structure, in an Aumann structure there is no function π that associates with each state in S a truth assignment to primitive propositions. (Using terminology we introduce in the next chapter, this means that an Aumann structure is really a *frame*.)

How do we define knowledge in the event-based approach? Since the objects of interest in this approach are events, it should not be surprising that knowledge is defined in terms of events. Formally, given an Aumann structure $(S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$, we define knowledge operators $\mathsf{K}_i : 2^S \to 2^S$, for $i = 1, \ldots, n$, as follows:

$$\mathsf{K}_i(e) = \{s \in S \mid \mathcal{P}_i(s) \subseteq e\};$$

$K_i(e)$ is called the event of i *knowing* e. Here 2^S is the set of all subsets of S. (Note that we use sans serif font for the knowledge operator K_i, in contrast to the italic font that we use for the modal operator K_i, and the script font we use for the binary relation $\mathcal{K}_i$.) It is easy to see that $K_i(e)$ is the union of the information sets of agent i that are contained in e. The intuition is that agent i knows e at state s if e holds at every state that agent i considers possible at state s (namely, at all states of $\mathcal{P}_i(s)$). Thus, agent i knows that no matter what the actual state is, the event e holds there.

The event of *everyone in a group G knowing* e is captured by an operator $E_G : 2^S \rightarrow 2^S$ defined as follows:

$$E_G(e) = \bigcap_{i \in G} K_i(e).$$

We can iterate the E_G operator, defining $E_G^1(e) = E_G(e)$ and $E_G^{k+1}(e) = E_G(E_G^k(e))$ for $k \geq 1$. *Common knowledge of an event e among the agents in a group G*, denoted $C_G(e)$, is the event of the players all knowing e, all knowing that all know it, and so on ad infinitum. Formally, we define

$$C_G(e) = \bigcap_{k=1}^{\infty} E_G^k(e).$$

Finally, *distributed knowledge of an event e among the agents in a group G*, denoted $D_G(e)$, is defined by

$$D_G(e) = \left\{ s \in S \mid \left(\bigcap_{i \in G} \mathcal{P}_i(s) \right) \subseteq e \right\}.$$

Intuitively, event e is distributed knowledge if e holds at all of the states that remain possible once we combine the information available to all of the agents.

Given two partitions $\mathcal{P}$ and $\mathcal{P}'$ of a set S, the partition $\mathcal{P}$ is said to be *finer* than $\mathcal{P}'$ (and $\mathcal{P}'$ to be *coarser* than $\mathcal{P}$) if $\mathcal{P}(s) \subseteq \mathcal{P}'(s)$ holds for all $s \in S$. Intuitively, if partition $\mathcal{P}$ is finer than partition $\mathcal{P}'$, then the information sets given by $\mathcal{P}$ give at least as much information as the information sets given by $\mathcal{P}'$ (since considering fewer states possible corresponds to having more information). The *meet* of partitions $\mathcal{P}$ and $\mathcal{P}'$, denoted $\mathcal{P} \sqcap \mathcal{P}'$, is the finest partition that is coarser than $\mathcal{P}$ and $\mathcal{P}'$; the *join* of $\mathcal{P}$ and $\mathcal{P}'$, denoted $\mathcal{P} \sqcup \mathcal{P}'$, is the coarsest partition finer than $\mathcal{P}$ and $\mathcal{P}'$. In the next proposition, we make use of the meet and the join to give nice characterizations of common knowledge and distributed knowledge.

Proposition 2.5.1 *Let $A = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$ be an Aumann structure, let $G \subseteq \{1, \ldots, n\}$ be a group of agents, and let $e \subseteq S$. Then*

(a) $s \in C_G(e)$ iff $(\sqcap_{i \in G} \mathcal{P}_i)(s) \subseteq e$.

(b) $s \in D_G(e)$ iff $(\sqcup_{i \in G} \mathcal{P}_i)(s) \subseteq e$.

Proof See Exercise 2.15. ∎

It follows that the meet of the agents' partitions characterizes their common knowledge, and the join of the agents' partitions characterizes their distributed knowledge. Notice that Proposition 2.5.1(a) implies that verifying whether an event e is common knowledge at a given state s can be done by one simple check of inclusion between two well-defined sets; it is unnecessary to use the definition of common knowledge, which involves an infinitary intersection.

There is a close connection between the logic-based approach and the event-based approach, which we now formalize. There is a natural one-to-one correspondence between partitions on S and equivalence relations on S. Given a partition $\mathcal{P}$ of S, the corresponding equivalence relation $\mathcal{R}$ is defined by $(s, s') \in \mathcal{R}$ iff $\mathcal{P}(s) = \mathcal{P}(s')$. Similarly, given an equivalence relation $\mathcal{R}$ on S, the corresponding partition $\{S_1, \ldots, S_r\}$ of S is obtained by making each equivalence class of $\mathcal{R}$ into a cell S_j of the partition; that is, two states s, t are in the same cell of the partition precisely if $(s, t) \in \mathcal{R}$. It is thus easy to convert back and forth between the partition viewpoint and the equivalence relations viewpoint (see Exercise 2.16).

Assume now that we are given a Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where each $\mathcal{K}_i$ is an equivalence relation. We define the corresponding Aumann structure $A^M = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$ (with the same set S of states) by taking $\mathcal{P}_i$ to be the partition corresponding to the equivalence relation $\mathcal{K}_i$. We want to show that M and A^M have the same "semantics." The semantics in M is defined in terms of formulas. The *intension* of a formula φ in structure M, denoted φ^M, is the set of states of M at which φ holds, i.e., $\varphi^M = \{s \mid (M, s) \models \varphi\}$. The semantics in A^M is defined in terms of events. For each primitive proposition p, define e_p^M to be the event that p is true; that is, $e_p^M = \{s \mid (M, s) \models p\}$. We can now define an event $\mathrm{ev}_M(\varphi)$ for each formula φ by induction on the structure of φ:

- $\mathrm{ev}_M(p) = e_p^M$

- $\mathrm{ev}_M(\psi_1 \wedge \psi_2) = \mathrm{ev}_M(\psi_1) \cap \mathrm{ev}_M(\psi_2)$

- $\mathrm{ev}_M(\neg \psi) = S - \mathrm{ev}_M(\psi)$

- $\mathrm{ev}_M(K_i \psi) = \mathsf{K}_i(\mathrm{ev}_M(\psi))$

- $\mathsf{ev}_M(C_G\psi) = \mathsf{C}_G(\mathsf{ev}_M(\psi))$

- $\mathsf{ev}_M(D_G\psi) = \mathsf{D}_G(\mathsf{ev}_M(\psi))$

Intuitively, $\mathsf{ev}_M(\varphi)$ is the event that φ holds. The following proposition shows that this intuition is correct, that is, that the formula φ holds at state s of the Kripke structure M iff $\mathsf{ev}_M(\varphi)$ holds at state s of the Aumann structure A^M.

Proposition 2.5.2 *Let M be a Kripke structure where each possibility relation $\mathcal{K}_i$ is an equivalence relation, and let A^M be the corresponding Aumann structure. Then for every formula φ, we have $\mathsf{ev}_M(\varphi) = \varphi^M$.*

Proof   See Exercise 2.17. ∎

We have just shown how to go from a Kripke structure to a corresponding Aumann structure. What about the other direction? Let $A = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$ be an Aumann structure. We want to define a corresponding Kripke structure $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ (with the same set S of states). Defining the $\mathcal{K}_i$'s is no problem: we simply take $\mathcal{K}_i$ to be the equivalence relation corresponding to the partition $\mathcal{P}_i$. What about the set Φ of primitive propositions and the function π that associates with each state in S a truth assignment to primitive propositions? Although an Aumann structure does not presuppose the existence of a set of primitive propositions, in concrete examples there typically are names for basic events of interest, such as "Alice wins the game" or "the deal is struck." These names can be viewed as primitive propositions. It is also usually clear at which states these named events hold; this gives us the function π. To formalize this, assume that we are given not only the Aumann structure A but also an arbitrary set Φ of primitive propositions and an arbitrary function π that associates with each state in S a truth assignment to primitive propositions in Φ. We can now easily construct a Kripke structure $M^{A,\pi}$, which corresponds to A and π. If $A = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$, then $M^{A,\pi} = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where $\mathcal{K}_i$ is the partition corresponding to $\mathcal{P}_i$, for $i = 1, \ldots, n$. It is straightforward to show that the Aumann structure corresponding to $M^{A,\pi}$ is A (see Exercise 2.18). Thus, by Proposition 2.5.2, the intensions of formulas in $M^{A,\pi}$ and the events corresponding to these formulas in A coincide.

Proposition 2.5.2 and the preceding discussion establish the close connection between the logic-based and event-based approaches that we claimed previously.

Exercises

2.1 Suppose it is common knowledge that all the children in the muddy children puzzle are blind. What would the graphical representation be of the Kripke structure describing the situation before the father speaks? What about after the father speaks?

* 2.2 Consider the following variant of the muddy children puzzle. Suppose that it is common knowledge that all the children except possibly child 1 are paying attention when the father speaks. Moreover, suppose the children have played this game with the father before, and it is common knowledge that when he speaks he says either "At least one of you has mud on your forehead" or a vacuous statement such as "My, this field is muddy." (Thus it is common knowledge that even if child 1 did not hear the father, he knows that the father made one of those statements.)

   (a) Describe the situation (i.e., the Kripke structure) after the father's statement. (Hint: each possible world can be characterized by an $(n + 2)$-tuple, where n is the total number of children.) Draw the Kripke structure for the case $n = 2$.

   (b) Can the children figure out whether or not they are muddy? (Hint: first consider the case where child 1 is not muddy, then consider the case where he is muddy and hears the father, and finally consider the case where he is muddy and does not hear the father.)

   (c) Can the children figure out whether or not they are muddy if the father says at the beginning "Two or more of you have mud on your forehead"?

2.3 Here is yet another variant of the muddy children puzzle. Suppose that the father says "Child number 1 has mud on his forehead" instead of saying "At least one of you has mud on your forehead." It should not be too hard to convince yourself that now the children (other than child 1) cannot deduce whether they have mud on their foreheads. Explain why this should be so (i.e., why the children cannot solve the puzzle in a situation where they apparently have *more* information). This example shows that another assumption inherent in the puzzle is that all relevant information has been stated in the puzzle, and in particular, that the father said no more than "At least one of you has mud on your forehead."

* 2.4 In this exercise, we formalize the aces and eights game from Exercise 1.1.

   (a) What are the possible worlds for this puzzle if the suit of the card matters? How many possible worlds are there?

(b) Now suppose we ignore the suit (so, for example, we do not distinguish a hand with the ace of clubs and the ace of hearts from a hand with the ace of spades and the ace of hearts). How many possible worlds are there in this case? Since the suit does not matter in the puzzle, we still get an adequate representation for the puzzle if we ignore it. Since there are so many fewer possible worlds to consider in this case, it is certainly a worthwhile thing to do.

(c) Draw the Kripke structure describing the puzzle.

(d) Consider the situation described in part (a) of Exercise 1.1. Which edges disappear from the structure when you hear that Alice and Bob cannot determine what cards they have? Explain why it is now crucial to think of edges, not nodes, as disappearing.

(e) Now consider the situation described in part (b) of Exercise 1.1 and show which edges disappear from the structure.

* **2.5** In this exercise, we formalize the wise men puzzle from Exercise 1.3.

(a) Consider the first version of the puzzle (as described in part (a) of Exercise 1.3). Draw the Kripke structure describing the initial situation. How does the structure change after the first wise man says that he does not know the color of the hat on his head? How does it change after the second wise man says that he does not know?

(b) How does the initial Kripke structure change if the third wise man is blind?

2.6 Show that G-reachability is an equivalence relation if the $\mathcal{K}_i$ relations are reflexive and symmetric.

2.7 Show that if t is G-reachable from s, then $(M, s) \models C_G \varphi$ iff $(M, t) \models C_G \varphi$, provided that the $\mathcal{K}_i$ relation is reflexive and symmetric.

2.8 Show that the following properties of common knowledge are valid, using semantic arguments as in Theorems 2.4.1 and 2.4.2:

(a) $(C_G \varphi \wedge C_G(\varphi \Rightarrow \psi)) \Rightarrow C_G \psi$,

(b) $C_G \varphi \Rightarrow \varphi$ (assuming that the $\mathcal{K}_i$ relations are reflexive),

(c) $C_G \varphi \Rightarrow C_G C_G \varphi$,

(d) $\neg C_G \varphi \Rightarrow C_G \neg C_G \varphi$ (assuming that the $\mathcal{K}_i$ relations are symmetric),

(e) $C_G \varphi \Rightarrow C_{G'} \varphi$ if $G \supseteq G'$.

As is shown in Exercise 3.11, these properties are actually provable from the properties of knowledge and common knowledge described in this chapter.

2.9 Show that if $M \models \varphi \Rightarrow \psi$, then

(a) $M \models K_i \varphi \Rightarrow K_i \psi$,

(b) $M \models C_G \varphi \Rightarrow C_G \psi$.

2.10 Show that the following properties of distributed knowledge are valid:

(a) $(D_G \varphi \wedge D_G(\varphi \Rightarrow \psi)) \Rightarrow D_G \psi$,

(b) $D_G \varphi \Rightarrow \varphi$ (assuming that the $\mathcal{K}_i$ relations are reflexive),

(c) $D_G \varphi \Rightarrow D_G D_G \varphi$ (assuming that the $\mathcal{K}_i$ relations are transitive),

(d) $\neg D_G \varphi \Rightarrow D_G \neg D_G \varphi$ (assuming that the $\mathcal{K}_i$ relations are symmetric and transitive),

(e) $D_{\{i\}} \varphi \Leftrightarrow K_i \varphi$,

(f) $D_G \varphi \Rightarrow D_{G'} \varphi$ if $G \subseteq G'$.

2.11 Prove using semantic arguments that knowledge and common knowledge distribute over conjunction; i.e., prove that the following properties are valid:

(a) $K_i(\varphi \wedge \psi) \Leftrightarrow (K_i \varphi \wedge K_i \psi)$,

(b) $C_G(\varphi \wedge \psi) \Leftrightarrow (C_G \varphi \wedge C_G \psi)$.

It can also be shown that these properties follow from the properties described for knowledge and common knowledge in the text (Exercise 3.31).

2.12 Prove that the following formulas are valid assuming that the $\mathcal{K}_i$ relations are equivalence relations:

(a) $\models \neg \varphi \Rightarrow K_i \neg K_i \varphi$,

(b) $\models \neg \varphi \Rightarrow K_{i_1} \ldots K_{i_k} \neg K_{i_k} \ldots K_{i_1} \varphi$ for any sequence $i_1, \ldots, i_k$ of agents,

(c) $\models \neg K_i \neg K_i \varphi \Leftrightarrow K_i \varphi$.

These formulas are also provable from the S5 properties we discussed; see Exercise 3.14.

2.13 Let $A = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$ be an Aumann structure, and let $G \subseteq \{1, \ldots, n\}$. If s, t are states, we say that t *is G-reachable from s in A* if t is reachable from s in a Kripke structure $M^{A,\pi}$ corresponding to A. Prove that $t \in (\sqcap_{i \in G} \mathcal{P}_i)(s)$ iff t is G-reachable from s.

2.14 Let $A = (S, \mathcal{P}_1, \ldots, \mathcal{P}_n)$ be an Aumann structure and let $G \subseteq \{1, \ldots, n\}$. Prove that $t \in (\sqcup_{i \in G} \mathcal{P}_i)(s)$ iff for every agent i we have $t \in \mathcal{P}_i(s)$.

2.15 Prove Proposition 2.5.1. (Hint: you may either prove this directly, or use Exercises 2.13 and 2.14.)

2.16 Show that the correspondence we have given between partitions and equivalence relations and the correspondence defined in the other direction are inverses. That is, show that $\mathcal{R}$ is the equivalence relation that we obtain from a partition $\mathcal{P}$ iff $\mathcal{P}$ is the partition that we obtain from the equivalence relation $\mathcal{R}$.

2.17 Let M be a Kripke structure where each possibility relation $\mathcal{K}_i$ is an equivalence relation, and let A be the corresponding Aumann structure.

(a) Prove that

  (i) $s \in K_i(\text{ev}(\varphi))$ holds in A iff $(M, s) \models K_i \varphi$,

  (ii) $s \in D_G(\text{ev}(\varphi))$ holds in A iff $(M, s) \models D_G \varphi$,

  (iii) $s \in C_G(\text{ev}(\varphi))$ holds in A iff $(M, s) \models C_G \varphi$.

(b) Use part (a) to prove Proposition 2.5.2.

2.18 Show that the Aumann structure corresponding to the Kripke structure $M^{A,\pi}$ is A.

Notes

Modal logic was discussed by several authors in ancient times, notably by Aristotle in *De Interpretatione* and *Prior Analytics*, and by medieval logicians, but like most work before the modern period, it was nonsymbolic and not particularly systematic in approach. The first symbolic and systematic approach to the subject appears to be the work of Lewis beginning in 1912 and culminating in the book *Symbolic Logic* with Langford [1959]. Carnap [1946, 1947] suggested using possible worlds to assign semantics to modalities. Possible-worlds semantics was further developed independently by several researchers, including Bayart [1958], Hintikka [1957, 1961], Kanger [1957b], Kripke [1959], Meredith [1956], Montague [1960], and Prior [1962] (who attributed the idea to P. T. Geach), reaching its current form (as presented here) with Kripke [1963a]. Many of these authors also observed that by varying the properties of the $\mathcal{K}_i$ relations, we can obtain different properties of knowledge.

The initial work on modal logic considered only the modalities of possibility and necessity. As we mentioned in the bibliographic notes of Chapter 1, the idea of capturing the semantics of knowledge in this way is due to Hintikka, who also first observed the properties of knowledge discussed in Section 2.4.

The analysis of the muddy children puzzle in terms of Kripke structures is due to Halpern and Vardi [1991a]. Aumann structures were defined by Aumann [1976]. Aumann defines common knowledge in terms of the meet; in particular, the observation made in Proposition 2.5.1(a) is due to Aumann. A related approach, also defining knowledge as an operator on events, is studied by Orlowska [1989].

state. This result has far-reaching implications in the design of communication protocols in current-day computer networks. For a formulation of this result in the terminology of Theorem 11.6.3, see the overview by Moses [1992].

Panangaden and Taylor [1992] define the notion of *concurrent common knowledge*, which is a variant of common knowledge that is achievable in reliable asynchronous systems. They discuss the extent to which it can be used in such systems as a substitute for common knowledge. A variant of common knowledge called *continual common knowledge* is defined by Halpern, Moses, and Waarts [1990] and used to analyze and find round-optimal algorithms for the *eventual* Byzantine agreement problem, which is defined just like simultaneous Byzantine agreement except that processes are not required to decide simultaneously (see Exercise 6.39). Continual common knowledge is the one variant of common knowledge that is typically strictly stronger than common knowledge, while the other variants of common knowledge discussed in the literature can be viewed as approximations of common knowledge.

Bibliography

Afrati, F., C. H. Papadimitriou, and G. Papageorgiou (1988). The synthesis of communication protocols. *Algorithmica 3*(3), 451–472.

Aho, A. V., J. D. Ullman, A. D. Wyner, and M. Yannakakis (1982). Bounds on the size and transmission rate of communication protocols. *Computers and Mathematics with Applications 8*(3), 205–214. This is a later version of [Aho, Ullman, and Yannakakis 1979].

Aho, A. V., J. D. Ullman, and M. Yannakakis (1979). Modeling communication protocols by automata. In *Proc. 20th IEEE Symp. on Foundations of Computer Science*, pp. 267–273.

Anderlini, L. (1989). Computability and communication in common interest games. mimeo, Cambridge University.

Anderson, A. and N. D. Belnap (1975). *Entailment: The Logic of Relevance and Necessity*. Princeton, N.J.: Princeton University Press.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics 4*(6), 1236–1239.

Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory 37*, 167–190.

Barcan, R. C. (1946). A functional calculus of first order based on strict implication. *Journal of Symbolic Logic 11*, 1–16.

Barcan, R. C. (1947). The identity of individuals in a strict functional calculus of second order. *Journal of Symbolic Logic 12*, 12–15.

Bartlett, K. A., R. A. Scantlebury, and P. T. Wilkinson (1969). A note on reliable full-duplex transmission over half-duplex links. *Communications of the ACM 12*, 260–261.

Barwise, J. (1981). Scenes and other situations. *Journal of Philosophy 78*(7), 369–397.

Barwise, J. (1988). Three views of common knowledge. In M. Y. Vardi (Ed.), *Proc. Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 365–379. San Francisco, Calif.: Morgan Kaufmann.

Barwise, J. and J. Perry (1983). *Situations and Attitudes*. Cambridge, Mass.: Bradford Books.

Bayart, A. (1958). La correction de la logique modale du premier et second ordre S5. *Logique et Analyse 1*, 28–44.

Bazzi, R. and G. Neiger (1992). The complexity and impossibility of achieving fault-tolerant coordination. In *Proc. 11th ACM Symp. on Principles of Distributed Computing*, pp. 203–214.

Belnap, N. D. (1977). A useful four-valued logic. In G. Epstein and J. M. Dunn (Eds.), *Modern Uses of Multiple-Valued Logic*, pp. 5–37. Dordrecht, Netherlands: Reidel.

Benthem, J. F. A. K. van (1974). Some correspondence results in modal logic. Report 74–05, University of Amsterdam.

Benthem, J. F. A. K. van (1983). *Modal Logic and Classical Logic*. Naples: Bibliopolis.

Berman, P., J. Garay, and K. J. Perry (1989). Towards optimal distributed consensus. In *Proc. 30th IEEE Symp. on Foundations of Computer Science*, pp. 410–415.

Białynicki-Birula, A. and H. Rasiowa (1957). On the representation of quasi-Boolean algebras. *Bull. de l'académie polonaise des sciences 5*, 259–261.

Binmore, K. (1987). Modeling rational players I. *Economics and philosophy 3*, 179–214. Part II appeared ibid., *4*, 9–55.

Binmore, K. (1990). *Essays on the Foundations of Game Theory*. Oxford, U.K.: Basil Blackwell.

Binmore, K. and H. S. Shin (1993). Algorithmic knowledge and game theory. In C. Bicchieri and M. L. D. Chiara (Eds.), *Knowledge, Belief, and Strategic Interaction*, Chapter 9. Cambridge, U.K.: Cambridge University Press.

Bochmann, G. V. and J. Gecsei (1977). A unified method for the specification and verification of protocols. In B. Gilchrist (Ed.), *Information Processing 77*, pp. 229–234. Amsterdam: North-Holland.

Boolos, G. (1979). *The Unprovability of Consistency*. Cambridge: Cambridge University Press.

Brafman, R., J.-C. Latombe, Y. Moses, and Y. Shoham (1994). Knowledge as a tool in motion planning under uncertainty. In R. Fagin (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pp. 208–224. San Francisco, Calif.: Morgan Kaufmann.

Brandenburger, A. and E. Dekel (1987). Common knowledge with probability 1. *Journal of Mathematical Economics 16*, 237–245.

Canning, D. (1992). Rationality, computability and Nash equilibrium. *Econometrica 60*, 877–888.

Carnap, R. (1946). Modalities and quantification. *Journal of Symbolic Logic 11*, 33–64.

Carnap, R. (1947). *Meaning and Necessity*. Chicago: University of Chicago Press.

Carver, M. (1989). Aces and eights. *Discover* (February), 88.

Cave, J. (1983). Learning to agree. *Economics Letters 12*, 147–152.

Chandy, K. M. and J. Misra (1986). How processes learn. *Distributed Computing 1*(1), 40–52.

Chandy, K. M. and J. Misra (1988). *Parallel Program Design: A Foundation*. Reading, Mass.: Addison-Wesley.

Chang, C. C. and H. J. Keisler (1990). *Model Theory* (3rd ed.). Amsterdam: North-Holland.

Chellas, B. F. (1980). *Modal Logic*. Cambridge, U.K.: Cambridge University Press.

Clark, H. H. and C. R. Marshall (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, and I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge, U.K.: Cambridge University Press.

Clarke, E. M., E. A. Emerson, and A. P. Sistla (1986). Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Trans. on Programming Languages and Systems 8*(2), 244–263. An early version appeared in *Proc. 10th ACM Symposium on Principles of Programming Languages*, 1983.

Conway, J. H., M. S. Paterson, and U. S. S. R. Moscow (1977). A headache-causing problem. In J. K. Lenstra et al. (Eds.), *Een pak met een korte broek: Papers presented to H. W. Lenstra on the occasion of the publication of his "Euclidische Getallenlichamen"*. Private publication.

Cook, S. A. (1971). The complexity of theorem proving procedures. In *Proc. 3rd ACM Symp. on Theory of Computing*, pp. 151–158.

Cresswell, M. J. (1970). Classical intensional logics. *Theoria 36*, 347–372.

Cresswell, M. J. (1972). Intensional logics and logical truth. *Journal of Philosophical Logic 1*, 2–15.

Cresswell, M. J. (1973). *Logics and Languages*. London: Methuen and Co.

DeMillo, R. A., N. A. Lynch, and M. J. Merritt (1982). Cryptographic protocols. In *Proc. 14th ACM Symp. on Theory of Computing*, pp. 383–400.

Diffie, W. and M. E. Hellman (1976). New directions in cryptography. *IEEE Transactions on Information Theory 22*(5), 644–654.

Dolev, D., J. Y. Halpern, and H. R. Strong (1986). On the possibility and impossibility of achieving clock synchronization. *Journal of Computer and System Sciences 32*(2), 230–250.

Dolev, D., R. Reischuk, and H. R. Strong (1990). Early stopping in Byzantine agreement. *Journal of the ACM 34*(7), 720–741.

Dolev, D. and H. R. Strong (1982). Polynomial algorithms for multiple processor agreement. In *Proc. 14th ACM Symp. on Theory of Computing*, pp. 401–407.

Dunn, J. M. (1976). Intuitive semantics for first-degree entailment and "coupled trees". *Philosophical Studies 29*, 149–168.

Dunn, J. M. (1986). Relevance logic and entailment. In D. Gabbay and F. Guenthner (Eds.), *Handbook of Philosophical Logic, Vol. III*, pp. 117–224. Dordrecht, Netherlands: Reidel.

Dwork, C. and Y. Moses (1990). Knowledge and common knowledge in a Byzantine environment: crash failures. *Information and Computation 88*(2), 156–186.

Eberle, R. A. (1974). A logic of believing, knowing and inferring. *Synthese 26*, 356–382.

Elgot-Drapkin, J. J. (1991). Step-logic and the three-wise-men problem. In *Proc. National Conference on Artificial Intelligence (AAAI '91)*, pp. 412–417.

Elgot-Drapkin, J. J. and D. Perlis (1990). Reasoning situation in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence 2*(1), 75–98.

Emde Boas, P. van, J. Groenendijk, and M. Stokhof (1980). The Conway paradox: its solution in an epistemic framework. In *Proc. 3rd Amsterdam Montague Symposium*, pp. 87–111. Math. Centrum, Amsterdam. Reprinted by Foris Publications, 1984.

Emerson, E. A. and J. Y. Halpern (1985). Decision procedures and expressiveness in the temporal logic of branching time. *Journal of Computer and System Sciences 30*(1), 1–24.

Emerson, E. A. and J. Y. Halpern (1986). "Sometimes" and "not never" revisited: on branching versus linear time temporal logic. *Journal of the ACM 33*(1), 151–178.

Enderton, H. B. (1972). *A Mathematical Introduction to Logic*. New York: Academic Press.

Fagin, R. and J. Y. Halpern (1988a). Belief, awareness, and limited reasoning. *Artificial Intelligence 34*, 39–76.

Fagin, R. and J. Y. Halpern (1988b). I'm OK if you're OK: on the notion of trusting communication. *Journal of Philosophical Logic 17*(4), 329–354. Reprinted in: R. H. Thomason (Ed.) (1989), *Philosophical Logic and Artificial Intelligence*, pp. 9–34, Dordrecht, Netherlands: Kluwer.

Fagin, R. and J. Y. Halpern (1994). Reasoning about knowledge and probability. *Journal of the ACM 41*(2), 340–367.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1994). Knowledge-based programming. Research Report RJ 9711, IBM.

Fagin, R., J. Y. Halpern, and M. Y. Vardi (1990). A nonstandard approach to the logical omniscience problem. In R. Parikh (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, pp. 41–55. San Francisco, Calif.: Morgan Kaufmann. To appear in *Artificial Intelligence*.

Fagin, R., J. Y. Halpern, and M. Y. Vardi (1992a). What can machines know? On the properties of knowledge in distributed systems. *Journal of the ACM 39*(2), 328–376.

Fagin, R., J. Y. Halpern, and M. Y. Vardi (1992b). What is an inference rule? *Journal of Symbolic Logic 57*(3), 1018–1045.

Fagin, R. and M. Y. Vardi (1986). Knowledge and implicit knowledge in a distributed environment: preliminary report. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 187–206. San Francisco, Calif.: Morgan Kaufmann.

Fischer, M. J. (1983). The consensus problem in unreliable distributed systems. Technical Report RR-273, Yale University.

Fischer, M. J. and N. Immerman (1986). Foundations of knowledge for distributed systems. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 171–186. San Francisco, Calif.: Morgan Kaufmann.

Fischer, M. J. and N. Immerman (1987). Interpreting logics of knowledge in propositional dynamic logic with converse. *Information Processing Letters 25*(3), 175–182.

Fischer, M. J. and R. E. Ladner (1979). Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences 18*(2), 194–211.

Fischer, M. J. and N. A. Lynch (1982). A lower bound on the time to assure interactive consistency. *Information Processing Letters 14*(4), 183–186.

Fischer, M. J. and L. D. Zuck (1987). Relative knowledge and belief (extended abstract). Technical Report YALEU/DCS/TR-589, Yale University.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik 100*, 25–50. English translation in P. Geach and M. Black (1952), On sense and reference, *Translations From the Writings of Gottlob Frege*, Oxford, U.K.: Basil Blackwell.

Friedman, N. and J. Y. Halpern (1994). A knowledge-based framework for belief change. Part I: Foundations. In R. Fagin (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pp. 44–64. San Francisco, Calif.: Morgan Kaufmann.

Fudenberg, D. and J. Tirole (1991). *Game Theory*. Cambridge, Mass.: MIT Press.

Gabbay, D., A. Pnueli, S. Shelah, and J. Stavi (1980). On the temporal analysis of fairness. In *Proc. 7th ACM Symp. on Principles of Programming Languages*, pp. 163–173.

Gamow, G. and M. Stern (1958). *Puzzle Math*. New York: Viking Press.

Gardner, M. (1977). The "jump proof" and its similarity to the toppling of a row of dominoes. *Scientific American 236*, 128–135.

Gardner, M. (1984). *Puzzles From Other Worlds*. New York: Viking Press.

Garson, J. W. (1977). Quantification in modal logic. In D. Gabbay and F. Guenthner (Eds.), *Handbook of Philosophical Logic, Vol. II*, pp. 249–307. Dordrecht, Netherlands: Reidel.

Geanakopolos, J. (1989). Game theory without partitions, and applications to speculation and consensus. Cowles Foundation Discussion Paper #914, Yale University.

Geanakopolos, J. (1992). Common knowledge. In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 255–315. San Francisco, Calif.: Morgan Kaufmann.

Geanaokoplos, J. and H. Polemarchakis (1982). We can't disagree forever. *Journal of Economic Theory 28*(1), 192–200.

Georgeff, M. P. (1985). Attributing attitudes to machines. Manuscript.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis 23*, 121–123.

Goldblatt, R. (1992). *Logics of Time and Computation* (2nd ed.). CSLI Lecture Notes Number 7. Center for Studies in Language and Information, Stanford University.

Goldwasser, S., S. Micali, and C. Rackoff (1989). The knowledge complexity of interactive proof systems. *SIAM Journal on Computing 18*(1), 186–208.

Gouda, M. (1985). On "A simple protocol whose proof isn't". *IEEE Transactions on Communications COM-33*(4), 382–384.

Gray, J. (1978). Notes on database operating systems. In R. Bayer, R. M. Graham, and G. Seegmuller (Eds.), *Operating Systems: An Advanced Course*, Lecture Notes in Computer Science, Vol. 66. Berlin/New York: Springer-Verlag. Also appears as IBM Research Report RJ 2188, 1978.

Grove, A. J. and J. Y. Halpern (1991). Naming and identity in a multi-agent epistemic logic. In J. A. Allen, R. Fikes, and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, pp. 301–312. San Francisco, Calif.: Morgan Kaufmann.

Hadzilacos, V. (1987). A knowledge-theoretic analysis of atomic commitment protocols. In *Proc. 6th ACM Symp. on Principles of Database Systems*, pp. 129–134. A revised version has been submitted for publication.

Hailpern, B. T. (1982). *Verifying Concurrent Processes using Temporal Logic*. Lecture Notes in Computer Science, Vol. 129. Berlin/New York: Springer-Verlag.

Hailpern, B. T. (1985). A simple protocol whose proof isn't. *IEEE Transactions on Communications COM-33*(4), 330–337.

Hailpern, B. T. and S. S. Owicki (1983). Modular verification of communication protocols. *IEEE Transactions on Communications COM-31*(1), 56–68.

Halpern, J. Y. (1986). Reasoning about knowledge: an overview. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 1–17. San Francisco, Calif.: Morgan Kaufmann. Reprinted in *Proc. National Computer Conference*, 1986, pp. 219–228.

Halpern, J. Y. (1987). Using reasoning about knowledge to analyze distributed systems. In J. F. Traub, B. J. Grosz, B. W. Lampson, and N. J. Nilsson (Eds.), *Annual Review of Computer Science, Vol. 2*, pp. 37–68. Palo Alto, Calif.: Annual Reviews Inc.

Halpern, J. Y. (1991). A note on knowledge-based protocols and specifications. Research Report RJ 8454, IBM.

Halpern, J. Y. (1993a). Reasoning about knowledge: a survey circa 1991. In A. Kent and J. G. Williams (Eds.), *Encyclopedia of Computer Science and Technology, Volume 27 (Supplement 12)*. New York: Marcel Dekker.

Halpern, J. Y. (1993b). Reasoning about only knowing with many agents. In *Proc. National Conference on Artificial Intelligence (AAAI '93)*, pp. 655–661.

Halpern, J. Y. and R. Fagin (1985). A formal model of knowledge, action, and communication in distributed systems: preliminary report. In *Proc. 4th ACM Symp. on Principles of Distributed Computing*, pp. 224–236.

Halpern, J. Y. and R. Fagin (1989). Modelling knowledge and action in distributed systems. *Distributed Computing 3*(4), 159–179. A preliminary version appeared in *Proc. 4th ACM Symposium on Principles of Distributed Computing*, 1985, with the title "A formal model of knowledge, action, and communication in distributed systems: Preliminary report".

Halpern, J. Y., N. Megiddo, and A. Munshi (1985). Optimal precision in the presence of uncertainty. *Journal of Complexity 1*, 170–196.

Halpern, J. Y. and Y. Moses (1984). Towards a theory of knowledge and ignorance. In *Proc. AAAI Workshop on Non-monotonic Logic*, pp. 125–143. Reprinted in K. Apt (Ed.), *Logics and Models of Concurrent Systems*, Springer-Verlag, Berlin/New York, pp. 459–476, 1985.

Halpern, J. Y. and Y. Moses (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM 37*(3), 549–587. A preliminary version appeared in *Proc. 3rd ACM Symposium on Principles of Distributed Computing*, 1984.

Halpern, J. Y. and Y. Moses (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence 54*, 319–379.

Halpern, J. Y., Y. Moses, and M. R. Tuttle (1988). A knowledge-based analysis of zero knowledge. In *Proc. 20th ACM Symp. on Theory of Computing*, pp. 132–147.

Halpern, J. Y., Y. Moses, and M. Y. Vardi (1994). Algorithmic knowledge. In R. Fagin (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pp. 255–266. San Francisco, Calif.: Morgan Kaufmann.

Halpern, J. Y., Y. Moses, and O. Waarts (1990). A characterization of eventual Byzantine agreement. In *Proc. 9th ACM Symp. on Principles of Distributed Computing*, pp. 333–346.

Halpern, J. Y. and J. H. Reif (1983). The propositional dynamic logic of deterministic, well-structured programs. *Theoretical Computer Science 27*, 127–165.

Halpern, J. Y. and M. R. Tuttle (1993). Knowledge, probability, and adversaries. *Journal of the ACM 40*(4), 917–962.

Halpern, J. Y. and M. Y. Vardi (1986). The complexity of reasoning about knowledge and time. In *Proc. 18th ACM Symp. on Theory of Computing*, pp. 304–315.

Halpern, J. Y. and M. Y. Vardi (1988a). The complexity of reasoning about knowledge and time in asynchronous systems. In *Proc. 20th ACM Symp. on Theory of Computing*, pp. 53–65.

Halpern, J. Y. and M. Y. Vardi (1988b). The complexity of reasoning about knowledge and time: synchronous systems. Research Report RJ 6097, IBM.

Halpern, J. Y. and M. Y. Vardi (1989). The complexity of reasoning about knowledge and time, I: lower bounds. *Journal of Computer and System Sciences 38*(1), 195–237.

Halpern, J. Y. and M. Y. Vardi (1991a). Model checking vs. theorem proving: a manifesto. In V. Lifschitz (Ed.), *Artificial Intelligence and Mathematical Theory of Computation (Papers in Honor of John McCarthy)*, pp. 151–176. San Diego, Calif.: Academic Press.

Halpern, J. Y. and M. Y. Vardi (1991b). Model checking vs. theorem proving: a manifesto. In J. A. Allen, R. Fikes, and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, pp. 325–334. San Francisco, Calif.: Morgan Kaufmann.

Halpern, J. Y. and L. D. Zuck (1992). A little knowledge goes a long way: knowledge-based derivations and correctness proofs for a family of protocols. *Journal of the ACM 39*(3), 449–478.

Heyting, A. (1956). *Intuitionism: An Introduction*. Amsterdam: North-Holland.

Hintikka, J. (1957). Quantifiers in deontic logic. *Societas Scientiarumi Fennica, Commentationes Humanarum Literarum 23*(4), 1–23.

Hintikka, J. (1961). Modalities and quantification. *Theoria 27*(61), 119–128.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic 4*, 475–484.

Hoare, C. A. R. (1985). *Communicating Sequential Processes*. Englewood Cliffs, N.J.: Prentice-Hall.

Hoek, W. van der and J.-J. C. Meyer (1992). Making some issues of implicit knowledge explicit. *Int'l. J. of Foundations of Computer Science 3*(2), 193–223.

Hopcroft, J. E. and J. D. Ullman (1979). *Introduction to Automata Theory, Languages and Computation*. New York: Addison-Wesley.

Huang, Z. and K. Kwast (1991). Awareness, negation and logical omniscience. In J. van Eijck (Ed.), *Logics in AI, Proceedings JELIA '90*, Lecture Notes in Computer Science, Vol. 478, pp. 282–300. Berlin/New York: Springer-Verlag.

Hughes, G. E. and M. J. Cresswell (1968). *An Introduction to Modal Logic*. London: Methuen.

Hughes, G. E. and M. J. Cresswell (1984). *A Companion to Modal Logic*. London: Methuen.

Kanger, S. (1957a). The morning star paradox. *Theoria 23*, 1–11.

Kanger, S. (1957b). *Provability in Logic*. Stockholm Studies in Philosophy I.

Kaplan, D. (1966). Review of "A semantical analysis of modal logic I: normal modal propositional calculi". *Journal of Symbolic Logic 31*, 120–122.

Kaplan, D. (1969). Quantifying in. *Synthese 19*, 178–214.

Katz, S. and G. Taubenfeld (1986). What processes know: definitions and proof methods. In *Proc. 5th ACM Symp. on Principles of Distributed Computing*, pp. 249–262.

Kifer, M. and E. L. Lozinski (1992). A logic for reasoning with inconsistency. *J. Automated Deduction 9*, 179–215.

Konolige, K. (1986). *A Deduction Model of Belief*. San Francisco, Calif.: Morgan Kaufmann.

Koo, R. and S. Toueg (1988). Effects of message loss on the termination of distributed programs. *Information Processing Letters 27*, 181–188.

Kozen, D. (1983). Results on the propositional μ-calculus. *Theoretical Computer Science 27*(1), 333–354.

Kozen, D. and R. Parikh (1981). An elementary proof of the completeness of PDL. *Theoretical Computer Science 14*(1), 113–118.

Krasucki, P., R. Parikh, and G. Ndjatou (1990). Probabilistic knowledge and probabilistic common knowledge (preliminary report). In Z. W. Ras, M. Zemankova, and M. L. Emrich (Eds.), *Methodologies for Intelligent Systems*, Volume 5, pp. 1–8. The Hague: Elsevier Science Publishing Co., Inc.

Kraus, S. and D. J. Lehmann (1988). Knowledge, belief, and time. *Theoretical Computer Science 58*, 155–174.

Kripke, S. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic 24*, 1–14.

Kripke, S. (1963a). A semantical analysis of modal logic I: normal modal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 9*, 67–96. Announced in (1959) *Journal of Symbolic Logic 24*, 323.

Kripke, S. (1963b). Semantical considerations on modal logic. *Acta Philosophica Fennica 24*, 83–94.

Kripke, S. (1965). A semantical analysis of modal logic II: non-normal propositional calculi. In L. Henkin and A. Tarski (Eds.), *The Theory of Models*, pp. 206–220. Amsterdam: North-Holland.

Kurki-Suonio, R. (1986). Towards programming with knowledge expressions. In *Proc. 13th ACM Symposium on Principles of Programming Languages*, pp. 140–149.

Ladner, R. E. (1977). The computational complexity of provability in systems of modal propositional logic. *SIAM Journal on Computing 6*(3), 467–480.

Ladner, R. E. and J. H. Reif (1986). The logic of distributed protocols (preliminary report). In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 207–222. San Francisco, Calif.: Morgan Kaufmann.

Lakemeyer, G. (1987). Tractable meta-reasoning in propositional logics of belief. In *Proc. Tenth International Joint Conference on Artificial Intelligence (IJCAI '87)*, pp. 402–408.

Lakemeyer, G. (1993). All they know: a study in multi-agent autoepestemic reasoning. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pp. 376–381.

Lamport, L. (1978). Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM 21*(7), 558–565.

Lamport, L. (1980). "Sometimes" is sometimes "not never": On the temporal logic of programs. In *Proc. 7th ACM Symp. on Principles of Programming Languages*, pp. 164–185.

Lamport, L. (1985). Paradigms for distributed computing. In M. Paul and H. J. Siegert (Eds.), *Methods and tools for specification, an advanced course*, Lecture Notes in Computer Science, Vol. 190, pp. 19–30, 454–468. Berlin/New York: Springer-Verlag.

Lamport, L. (1986). On interprocess communication, Part I: basic formalism. *Distributed Computing 1*(2), 77–85.

Lamport, L. and M. J. Fischer (1982). Byzantine generals and transactions commit protocols. Technical Report Opus 62, SRI International, Menlo Park, Calif.

Lehmann, D. J. (1984). Knowledge, common knowledge, and related puzzles. In *Proc. 3rd ACM Symp. on Principles of Distributed Computing*, pp. 62–67.

Lemmon, E. J. (1977). *The "Lemmon Notes": An Introduction to Modal Logic*. Oxford, U.K.: Basil Blackwell. Written in collaboration with Dana Scott; edited by Krister Segerberg. American Philosophical Quarterly Monograph Series. Monograph No. 11.

Lenzen, W. (1978). Recent work in epistemic logic. *Acta Philosophica Fennica 30*, 1–219.

Levesque, H. J. (1981). The interaction with incomplete knowledge bases: a formal treatment. In *Proc. Seventh International Joint Conference on Artificial Intelligence (IJCAI '81)*, pp. 240–245.

Levesque, H. J. (1984a). Foundations of a functional approach to knowledge representation. *Artificial Intelligence 23*, 155–212.

Levesque, H. J. (1984b). A logic of implicit and explicit belief. In *Proc. National Conference on Artificial Intelligence (AAAI '84)*, pp. 198–202.

Levesque, H. J. (1985). Global and local consistency and completeness of beliefs. Manuscript.

Levesque, H. J. (1988). Logic and the complexity of reasoning. *Journal of Philosophical Logic 17*(4), 355–389.

Levesque, H. J. (1990). All I know: A study in autoepistemic logic. *Artificial Intelligence 42*(3), 263–309.

Lewis, C. I. and C. H. Langford (1959). *Symbolic Logic* (2nd ed.). New York: Dover.

Lewis, D. (1969). *Convention, A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Lichtenstein, O., A. Pnueli, and L. Zuck (1985). The glory of the past. In R. Parikh (Ed.), *Proc. Workshop on Logics of Programs*, Lecture Notes in Computer Science, Vol. 193, pp. 196–218. Berlin/New York: Springer-Verlag.

Lipman, B. L. (1992a). Decision theory with impossible possible worlds. Technical Report Working paper, Queen's University.

Lipman, B. L. (1992b). Logics for nonomniscient agents: An axiomatic approach. Technical Report Working paper 874, Queen's University. Forthcoming in *Proceedings of the Second Castglioncello Conference*, C. Bicchieri and B. Skyrms (editors), Cambridge University Press.

Lipman, B. L. (1994). An axiomatic approach to the logical omniscience problem. In R. Fagin (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pp. 182–196. San Francisco, Calif.: Morgan Kaufmann.

Lynch, N. A. and M. J. Fischer (1981). On describing the behavior and implementation of distributed systems. *Theoretical Computer Science 13*, 17–43.

Lynch, N. A. and M. R. Tuttle (1989). An introduction to input/output automata. *CWI Quarterly 2*(3), 219–246. Also available as MIT Technical Memo MIT/LCS/TM-373.

Makinson, D. (1966). On some completeness theorems in modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 12*, 379–384.

Manna, Z. and A. Pnueli (1992). *The Temporal Logic of Reactive and Concurrent Systems*, Volume 1. Berlin/New York: Springer-Verlag.

Mazer, M. S. (1990). A link between knowledge and communication in faulty distributed systems. In R. Parikh (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, pp. 289–304. San Francisco, Calif.: Morgan Kaufmann.

Mazer, M. S. (1991). Implementing distributed knowledge-based protocols. Submitted for publication.

Mazer, M. S. and F. H. Lochovsky (1990). Analyzing distributed commitment by reasoning about knowledge. Technical Report CRL 90/10, DEC-CRL.

McCarthy, J. (1979). Ascribing mental qualities to machines. Technical Report STAN-CS-79-725, Stanford University.

McCarthy, J., M. Sato, T. Hayashi, and S. Igarishi (1979). On the model theory of knowledge. Technical Report STAN-CS-78-657, Stanford University.

Megiddo, N. (1986). Remarks on bounded rationality. Research Report RJ 5270, IBM.

Megiddo, N. (1989). On computable beliefs of rational machines. *Games and Economical Behavior 1*, 144–169.

Megiddo, N. and A. Wigderson (1986). On play by means of computing machines. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 259–274. San Francisco, Calif.: Morgan Kaufmann.

Meredith, C. A. (1956). Interpretations of different modal logics in the "property calculus", mimeographed manuscript, Philosophy Department, Canterbury University College (recorded and expanded by A. N. Prior).

Merritt, M. J. (1984). Unpublished notes on the Dolev-Strong lower bound for Byzantine agreement.

Merritt, M. J. and G. Taubenfeld (1991). Knowledge in shared memory systems. In *Proc. 10th ACM Symp. on Principles of Distributed Computing*, pp. 189–200.

Meyden, R. van der (1994). Axioms for knowledge and time in distributed systems with perfect recall. In *Proc. 9th IEEE Symp. on Logic in Computer Science*, pp. 448–457.

Meyer, J.-J. C., W. van der Hoek, and G. A. W. Vreeswijk (1991a). Epistemic logic for computer science: a tutorial (Part I). *EATCS Bulletin 44*, 242–270.

Meyer, J.-J. C., W. van der Hoek, and G. A. W. Vreeswijk (1991b). Epistemic logic for computer science: a tutorial (Part II). *EATCS Bulletin 45*, 256–287.

Michel, R. (1989a). A categorical approach to distributed systems, expressibility and knowledge. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pp. 129–143.

Michel, R. (1989b). *Knowledge in Distributed Byzantine Environments*. Ph. D. thesis, Yale University.

Milgrom, P. (1981). An axiomatic characterization of common knowledge. *Econometrica 49*(1), 219–222.

Milgrom, P. and N. Stokey (1982). Information, trade, and common knowledge. *Journal of Economic Theory 26*, 17–27.

Milner, R. (1980). *A Calculus of Communicating Systems*. Lecture Notes in Computer Science, Vol. 92. Berlin/New York: Springer-Verlag.

Monderer, D. and D. Samet (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior 1*, 170–190.

Montague, R. (1960). Logical necessity, physical necessity, ethics, and quantifiers. *Inquiry 4*, 259–269.

Montague, R. (1968). Pragmatics. In R. Kalibansky (Ed.), *Contemporary Philosophy*, pp. 101–121. Florence, Italy: La Nuova Italia Editrice.

Montague, R. (1970). Universal grammar. *Theoria 36*, 373–398.

Moore, R. C. (1985). A formal theory of knowledge and action. In J. Hobbs and R. C. Moore (Eds.), *Formal Theories of the Commonsense World*, pp. 319–358. Norwood, N.J.: Ablex Publishing Corp.

Moore, R. C. and G. Hendrix (1979). Computational models of beliefs and the semantics of belief sentences. Technical Note 187, SRI International, Menlo Park, Calif.

Moses, Y. (1988). Resource-bounded knowledge. In M. Y. Vardi (Ed.), *Proc. Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 261–276. San Francisco, Calif.: Morgan Kaufmann.

Moses, Y. (1992). Knowledge and communication (a tutorial). In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 1–14. San Francisco, Calif.: Morgan Kaufmann.

Moses, Y. and B. Bloom (1994). Knowledge, timed precedence and clocks. In *Proc. 13th ACM Symp. on Principles of Distributed Computing*, pp. 294–303.

Moses, Y., D. Dolev, and J. Y. Halpern (1986). Cheating husbands and other stories: a case study of knowledge, action, and communication. *Distributed Computing 1*(3), 167–176.

Moses, Y. and O. Kislev (1993). Knowledge-oriented programming. In *Proc. 12th ACM Symp. on Principles of Distributed Computing*, pp. 261–270.

Moses, Y. and G. Nachum (1990). Agreeing to disagree after all. In R. Parikh (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, San Francisco, Calif., pp. 151–168. Morgan Kaufmann.

Moses, Y. and G. Roth (1989). On reliable message diffusion. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pp. 119–128.

Moses, Y. and Y. Shoham (1993). Belief as defeasible knowledge. *Artificial Intelligence 64*(2), 299–322.

Moses, Y. and M. R. Tuttle (1988). Programming simultaneous actions using common knowledge. *Algorithmica 3*, 121–169.

Neiger, G. (1988). Knowledge consistency: a useful suspension of disbelief. In M. Y. Vardi (Ed.), *Proc. Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 295–308. San Francisco, Calif.: Morgan Kaufmann.

Neiger, G. and R. Bazzi (1992). Using knowledge to optimally achieve coordination in distributed systems. In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 43–59. San Francisco, Calif.: Morgan Kaufmann.

Neiger, G. and S. Toueg (1990). Automatically increasing the fault-tolerance of distributed algorithms. *Journal of Algorithms 11*(3), 374–419.

Neiger, G. and S. Toueg (1993). Simulating real-time clocks and common knowledge in distributed systems. *Journal of the ACM 40*(2), 334–367.

Neiger, G. and M. R. Tuttle (1993). Common knowledge and consistent simultaneous coordination. *Distributed Computing 6*(3), 334–352.

Newell, A. (1982). The knowledge level. *Artificial Intelligence 18*, 87–127.

Neyman, A. (1985). Bounded complexity justifies cooperation in finitely repated prisoner's dilemma. *Economic Letters*, 227–229.

Orlowska, E. (1989). Logic for reasoning about knowledge. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 35*, 559–572.

Panangaden, P. and S. Taylor (1992). Concurrent common knowledge: Defining agreement for asynchronous systems. *Distributed Computing 6*(2), 73–94.

Parikh, R. (1987). Knowledge and the problem of logical omniscience. In Z. W. Ras and M. Zemankova (Eds.), *Methodologies of Intelligent Systems*, pp. 432–439. The Hague: Elsevier Science Publishing Co., Inc.

Parikh, R. (1990). Recent issues in reasoning about knowledge. In R. Parikh (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, pp. 3–10. San Francisco, Calif.: Morgan Kaufmann.

Parikh, R. (1991). Monotonic and nonmonotonic logics of knowledge. *Fundamenta Informaticae 15*(3,4), 255–274.

Parikh, R. (1992). Finite and infinite dialogues. In Y. N. Moshovakis (Ed.), *Logic from Computer Science, MSRI Publication No. 21*, pp. 481–497. Berlin/New York: Springer-Verlag.

Parikh, R. and P. Krasucki (1990). Communication, consensus, and knowledge. *Journal of Economic Theory 52*(1), 178–189.

Parikh, R. and P. Krasucki (1992). Levels of knowledge in distributed computing. *Sādhanā 17*(1), 167–191.

Parikh, R. and R. Ramanujam (1985). Distributed processing and the logic of knowledge. In R. Parikh (Ed.), *Proc. Workshop on Logics of Programs*, pp. 256–268.

Patel-Schneider, P. F. (1985). A decidable first-order logic for knoweledge representation. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI '85)*, pp. 455–458.

Patel-Schneider, P. F. (1989). A four-valued semantics for terminological logics. *Artificial Intelligence 38*, 319–351.

Pease, M., R. Shostak, and L. Lamport (1980). Reaching agreement in the presence of faults. *Journal of the ACM 27*(2), 228–234.

Perrault, C. R. and P. R. Cohen (1981). It's for your own good: a note on inaccurate reference. In A. K. Johsi, B. L. Webber, and I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge, U.K.: Cambridge University Press.

Plantinga, A. (1974). *The Nature of Necessity*. Oxford, U.K.: Oxford University Press.

Pratt, V. R. (1979). Models of program logics. In *Proc. 20th IEEE Symp. on Foundations of Computer Science*, pp. 115–122.

Pratt, V. R. (1982). On the composition of processes. In *Proc. 9th ACM Symp. on Principles of Programming Languages*, pp. 213–223.

Pratt, V. R. (1985). Modelling concurrency with partial orders. *International Journal of Parallel Programming 15*(1), 33–71.

Prior, A. N. (1956). Modality and quantification in S5. *Journal of Symbolic Logic 21*, 60–62.

Prior, A. N. (1957). *Time and Modality.* Oxford, U.K.: Oxford University Press.

Prior, A. N. (1962). Possible worlds (idea attributed to P. T. Geach). *Philosophical Quarterly 12*, 36–43.

Quine, W. V. O. (1947). The problem of interpreting modal logic. *Journal of Symbolic Logic 12*, 43–48.

Rantala, V. (1982). Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica 35*, 18–24.

Rasmusen, E. (1989). *Games and Information: An Introduction to Game Theory.* Oxford, U.K. and Cambridge, Mass.: Basil Blackwell.

Rescher, N. and R. Brandom (1979). *The Logic of Inconsistency.* Totowa, N.J.: Rowman and Littlefield.

Rivest, R. L., A. Shamir, and L. Adelman (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM 21*(2), 120–126.

Rogers, Jr., H. (1967). *Theory of Recursive Functions and Effective Computability.* New York: McGraw-Hill.

Rosenschein, S. J. (1985). Formal theories of AI in knowledge and robotics. *New Generation Computing 3*, 345–357.

Rosenschein, S. J. and L. P. Kaelbling (1986). The synthesis of digital machines with provable epistemic properties. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 83–97. San Francisco, Calif.: Morgan Kaufmann.

Roth, G. (1989). Message diffusion in anonymous distributed systems. Master's thesis, Weizmann Institute of Science.

Routley, R. and V. Routley (1972). Semantics of first degree entailment. *Noûs 6*, 335–359.

Rubinstein, A. (1985). Finite automata play the repeated prisoner's dilemma. ST/ICERD Discussion Paper 85/109, London School of Economics.

Rubinstein, A. (1989). The electronic mail game: strategic behavior under "almost common knowledge". *American Economic Review 79*, 385–391.

Rubinstein, A. and A. Wolinsky (1990). On the logic of "agreeing to disagree" type results. *Journal of Economic Theory 51*, 184–193.

Samet, D. (1987). Ignoring ignorance and agreeing to disagree. MEDS discussion paper, Northwestern University.

Sanders, B. (1991). A predicate transformer approach to knowledge and knowledge-based protocols. In *Proc. 10th ACM Symp. on Principles of Distributed Computing*, pp. 217–230. A revised report appears as ETH Informatik Technical Report 181, 1992.

Savage, L. J. (1954). *Foundations of Statistics.* New York: John Wiley & Sons.

Scott, D. (1970). Advice on modal logic. In K. Lambert (Ed.), *Philosophical Problems in Logic*, pp. 143–173. Dordrecht, Netherlands: Reidel.

Segerberg, K. (1968). *Results in Nonclassical Logic.* Lund, Sweden: Berlingska Boktryckeriet.

Segerberg, K. (1971). *An Essay in Classical Modal Logic.* Philosophical Studies 13. Uppsala University.

Shalpey, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences 39*, 1095–1100.

Shin, H. S. and T. Williamson (1993). Representing the knowledge of Turing machines. Manuscript, University College, Oxford.

Shoham, Y. (1993). Agent oriented programming. *Artificial Intelligence 60*(1), 51–92.

Sistla, A. P. and E. M. Clarke (1985). The complexity of propositional linear temporal logics. *Journal of the ACM 32*(3), 733–749.

Spaan, E. (1990). Nexttime is not necessary. In R. J. Parikh (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Third Conference*, pp. 241–256. San Francisco, Calif.: Morgan Kaufmann.

Stalnaker, R. (1985). *Inquiry.* Cambridge, Mass.: MIT Press.

Stark, W. R. (1981). A logic of knowledge. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 27*, 371–374.

Stenning, M. V. (1976). A data transfer protocol. *Comput. Networks 1*, 99–110.

Tarski, A. (1955). A lattice-theoretic fixpoint theorem and its applications. *Pacific Journal of Mathematics 5*, 285–309.

Thijsse, E. (1992). *On Partial Logic and Knowledge Representation.* Netherlands: Eburon, Delft. (Ph.D. dissertation, Tilburg University.).

Thijsse, E. (1993). On total awareness logics. In M. de Rijke (Ed.), *Diamonds and Defaults*, pp. 309–347. Dordrecht, Netherlands: Kluwer.

Thomason, R. H. (1984). Combinations of tense and modality. In D. Gabbay and F. Guenthner (Eds.), *Handbook of Philosophical Logic, Vol. II*, pp. 135–165. Dordrecht, Netherlands: Reidel.

Thorp, E. O. (1961). A favorable strategy for twenty-one. *Proc. Natl. Acad. Sci. 47*(1), 110–112.

Thorp, E. O. (1966). *Beat the dealer* (2nd ed.). New York: Vintage.

Vardi, M. Y. (1985). A model-theoretic analysis of monotonic knowledge. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI '85)*, pp. 509–512.

Vardi, M. Y. (1989). On the complexity of epistemic reasoning. In *Proc. 4th IEEE Symp. on Logic in Computer Science*, pp. 243–252.

Voorbraak, F. (1992). Generalized Kripke models for epistemic logic. In Y. O. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 214–228. San Francisco, Calif.: Morgan Kaufmann.

Wansing, H. (1990). A general possible worlds framework for reasoning about knowledge and belief. *Studia Logica 49*(4), 523–539.

Wright, G. H. von (1951). *An Essay in Modal Logic*. Amsterdam: North-Holland.

Yemini, Y. and D. Cohen (1979). Some issues in distributed processes communication. In *Proc. of the 1st International Conf. on Distributed Computing Systems*, pp. 199–203.

Zadrozny, W. (1985). Explicit and implicit beliefs, a solution of a problem of H. Levesque. Manuscript.

Index