Reasoning about Knowledge

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

The MIT Press
Cambridge, Massachusetts
London, England

Chapter 1

Introduction and Overview

An investment in knowledge pays the best interest.

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

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

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

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

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

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

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

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

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

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

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

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

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

1.1 The "Muddy Children" Puzzle

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

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

Let us denote the fact "at least one child has a muddy forehead" by p. Notice that if $k > 1$, that is, 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, that is, 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 $(q+1)^{\text{st}}$ 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 children hear the answer to the $(k-1)^{\text{st}}$ question that the ones 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

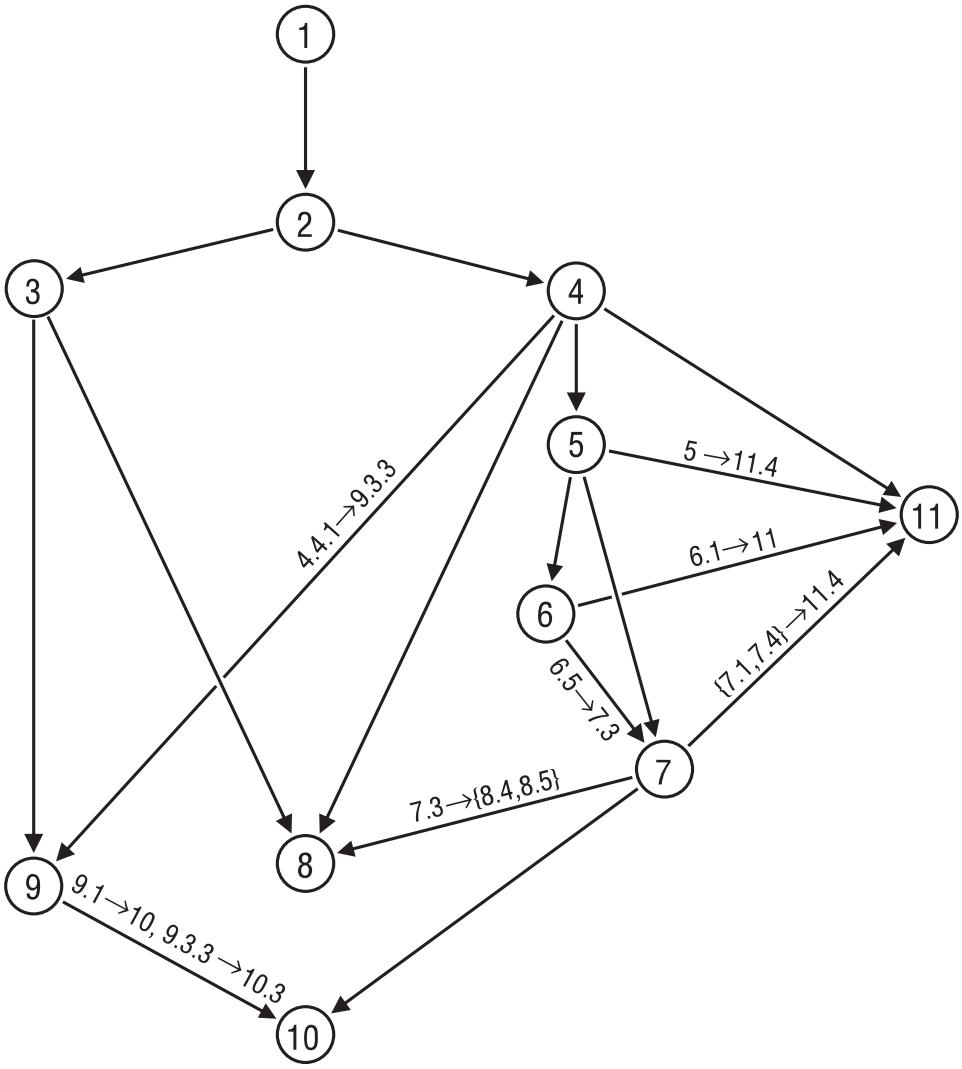

Figure 1.1 Dependency diagram

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 paradox in Chapter 11 and suggest two possible resolutions. The first makes use of the observation that if we model time at a sufficiently coarse level of granularity, then we often can and do attain common knowledge. The question then becomes when and whether it is appropriate to model time in this way. The second involves considering close approximations of common knowledge that are often attainable, and suffice for our purposes.

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 that 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 1960's and 1970's. 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 works by Halpern [1986, 1987,

1995], by Meyer, van der Hoek, and Vreeswijk [1991a, 1991b] (see also [Meyer and Hoek 1995]), 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 discussed first, in an informal way, by Hayek [1945], and then, in a more formal way, by Hilpinen [1977]. It was rediscovered and popularized 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 Littlewood [1953] and 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 game 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]. The wise men puzzle discussed in Exercise 1.3 seems to have been first discussed formally by McCarthy [1978], although it is undoubtedly much older. The well-known *surprise test paradox*, also known as the *surprise examination paradox*, the *hangman's paradox*, or the *unexpected hanging paradox*, is quite different from the wise men puzzle, but it too can be analyzed in terms of knowledge. Binkley [1968] does an analysis that explicitly uses knowledge; Chow [1998] gives a more up-to-date discussion. Halpern and Moses [1986] give a slightly different logic-based analysis, as well as pointers to the literature.

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, 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 that 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 herself 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, upon 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 that we have a group consisting of n agents, creatively named $1, \ldots, n$. For simplicity, we assume that 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

$$K_1K_2p \wedge \neg K_2K_1K_2p$$

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

$$\neg K_1\neg(K_2K_1K_2p) \wedge \neg K_1\neg(\neg K_2K_1K_2p).$$

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 above, 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 nonempty set of *states* or *possible worlds*, π is an *interpretation* which associates with each state in S a truth assignment to the primitive propositions in Φ (i.e., $\pi(s) : \Phi \rightarrow \{\mathbf{true}, \mathbf{false}\}$ 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.

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 $\pi(s)(p) = \mathbf{true}$ 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 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 $\pi(s)(p) = \mathbf{true}$.

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 that $\Phi = \{p\}$ 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 $\pi(s)(p) = \pi(u)(p) = \mathbf{true}$ and $\pi(t)(p) = \mathbf{false}$), 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.

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

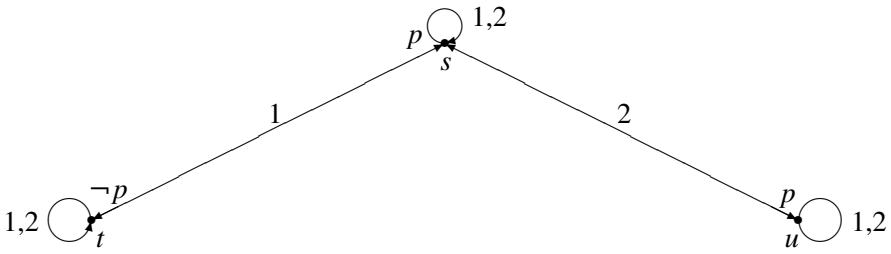

Figure 2.1 A simple Kripke structure

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. (And so, assuming that agent 2 were reliable, agent 1 knows that he could find out the true situation by asking agent 2.) 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:

$$(M, s) \models p \wedge \neg K_1 p \wedge K_2 p \wedge K_1(K_2 p \vee K_2 \neg p) \wedge \neg K_2 \neg K_1 p.$$

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 that 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 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, while 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 he has, namely, the card he is holding. This is an important general phenomenon: in any situation where an agent's possibility relation is determined by his 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.)

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 an 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).

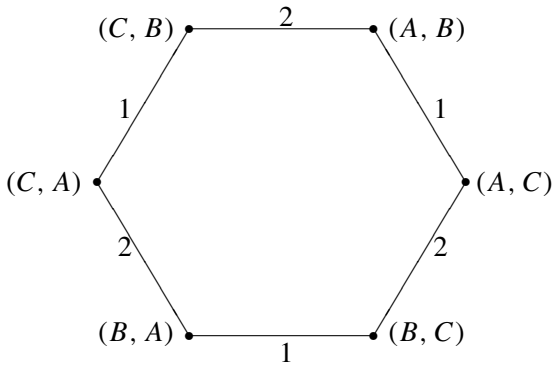

Figure 2.2 The Kripke structure describing a simple card game

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 an A, then she 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 an 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 there do not exist other reasonable notions of knowledge. 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 restrict to 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 own forehead, and that all the children are truthful and (extremely) intelligent.

First consider the situation before the father speaks. Suppose that there are n children altogether. As before, we number them $1, \ldots, n$. Some of the children have muddy foreheads, while 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 that 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 himself, his only doubt is about whether he has mud on his 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 i^{th} 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", while 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 i^{th} component, we take $(s, t) \in \mathcal{K}_i$ exactly if s and t agree in all components except possibly the i^{th} component. Notice that this definition makes $\mathcal{K}_i$ an equivalence relation. This completes the description of M.

While this Kripke structure may seem quite complicated, it actually has an elegant graphical representation. Suppose that 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 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 his own forehead is muddy, since in the other world he considers possible—(0,0,1)—his 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 $p_2 \Rightarrow K_1 p_2$,

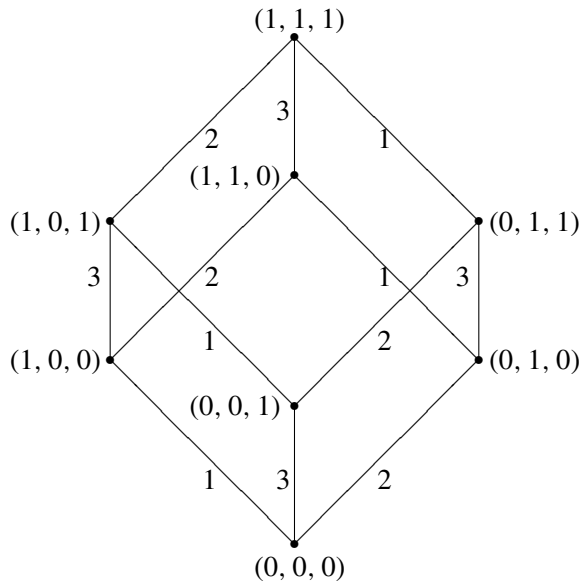

Figure 2.3 The Kripke structure for the muddy children puzzle with $n = 3$

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 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. Going back to our example with $n = 3$, 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.

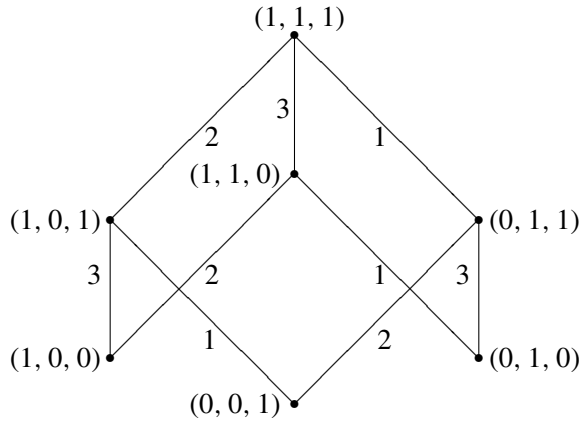

Figure 2.4 The Kripke structure after the father speaks

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 parallels that done 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 $(0, 0, \ldots, 0)$. Since once the father speaks it is common knowledge that $(0, 0, \ldots, 0)$ is not possible, 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 $k+1$ 1's, 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 (it is common knowledge that) all reasoners are *logically omniscient*, that is, that they are smart enough to compute all the consequences of the information that they have, and that this logical omniscience is common knowledge.

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 "Does 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 does not have mud on his 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 m^{th} 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*

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

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 he 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*, while 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 that $(M, s) \models K_i\varphi$. 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 $(M, s) \models K_iK_i\varphi$.

(e) Suppose that $(M, s) \models \neg K_i\varphi$. Then for some u with $(s, u) \in \mathcal{K}_i$, we must have $(M, u) \models \neg\varphi$. Suppose that 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 $(M, s) \models K_i\neg K_i\varphi$. ∎

The collection of properties that we have considered so far—the Distribution Axiom, the Knowledge Axiom, 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. While taking $\mathcal{K}_i$ to be an equivalence relation seems reasonable for many applications we have in mind, one 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. While, of course, a number of additional properties follow from the basic S5 properties defined above, 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, two properties that 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:

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

Thus, the Fixed-Point Axiom says that $C_G\varphi$ can be viewed as a *fixed point* of the function $f(x) = E_G(\varphi \wedge x)$, which maps a formula x to the formula $E_G(\varphi \wedge x)$. (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 $M \models \varphi \Rightarrow E_G(\psi \wedge \varphi)$, then $M \models \varphi \Rightarrow C_G\psi$.

This rule is often called the *Induction Rule*inference rule!RC1 (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 that $(M, s) \models C_G\varphi$. 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 that $(M, s) \models E_G(\varphi \wedge C_G\varphi)$.

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, $(M, t) \models \varphi$ using Lemma 2.2.1 and the fact that $(M, s') \models C_G\varphi$. Since $(M, t) \models \varphi$ for all t that are G-reachable from s, it follows that $(M, s) \models C_G\varphi$.

Finally, for part (c), suppose that $M \models \varphi \Rightarrow E_G(\psi \wedge \varphi)$ 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 that 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 $(M, s) \models C_G\psi$. ∎

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 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.

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 that 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 that 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 (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". However, 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** (A formalization of 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 that 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?

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

*** 2.5** (A formalization of 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 all 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) $C_G\varphi \Rightarrow \varphi$,

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

(d) $\neg C_G\varphi \Rightarrow C_G\neg C_G\varphi$,

(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 all valid:

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

(b) $D_G\varphi \Rightarrow \varphi$,

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

(d) $\neg D_G\varphi \Rightarrow D_G\neg D_G\varphi$,

(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; that is, 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:

(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 and 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) $s \in \mathsf{K}_i(\mathtt{ev}(\varphi))$ holds in A iff $(M, s) \models K_i\varphi$,

(ii) $s \in \mathsf{D}_G(\mathtt{ev}(\varphi))$ holds in A iff $(M, s) \models D_G\varphi$,

(iii) $s \in \mathsf{C}_G(\mathtt{ev}(\varphi))$ 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 [1991]. 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]. Yet another approach, pursued in [Brandenburger and Dekel 1993; Emde Boas, Groenendijk, and Stokhof 1980; Fagin, Geanakoplos, Halpern, and Vardi 1999; Fagin, Halpern, and Vardi 1991; Fagin and Vardi 1985; Heifetz and Samet 1999; Mertens and Zamir 1985], models knowledge directly, rather than in terms of possible worlds. The key idea there is the construction of an infinite hierarchy of knowledge levels. The relation between that approach and the possible-world approach is discussed in [Fagin, Halpern, and Vardi 1991].

Chapter 3

Completeness and Complexity

There are four sorts of men:
He who knows not and knows not he knows not: he is a fool—shun him;
He who knows not and knows he knows not: he is simple—teach him;
He who knows and knows not he knows: he is asleep—wake him;
He who knows and knows he knows: he is wise—follow him.

Arabian proverb

In Chapter 2 we discussed the properties of knowledge (as well as of common knowledge and distributed knowledge). We attempted to characterize these properties in terms of valid formulas. All we did, however, was to list *some* valid properties. It is quite conceivable that there are additional properties of knowledge that are not consequences of the properties listed in Chapter 2. In this chapter, we give a complete characterization of the properties of knowledge. 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 listed in Chapter 2. The second approach is *algorithmic*: we study algorithms that recognize the valid properties of knowledge. We also consider the computational complexity of recognizing valid properties of knowledge. Doing so will give us some insight into what makes reasoning about knowledge difficult.

When analyzing the properties of knowledge, it is useful to consider a somewhat more general framework than that of the previous chapter. Rather than restrict attention to the case where the possibility relations (the $\mathcal{K}_i$'s) are equivalence relations, we consider other binary relations as well. Although our examples show that taking

the $\mathcal{K}_i$'s to be equivalence relations is reasonably well-motivated, particularly when what an agent considers possible is determined by his information, there are certainly other choices possible. The real question is what we mean by "in world s, agent i considers world t possible."

Let us now consider an example where reflexivity might not hold. We can easily imagine an agent who refuses to consider certain situations possible, even when they are not ruled out by his information. Thus, Fred might refuse to consider it possible that his son Harry is taking illegal drugs, even if Harry is. Fred might claim to "know" that Harry is drug-free, since in all worlds Fred considers possible, Harry is indeed drug-free. In that case, Fred's possibility relation would not be reflexive; in world s where Harry is taking drugs, Fred would not consider world s possible. To see why symmetry might not hold, consider poor Fred again. Suppose that in world s, Fred's wife Harriet is out visiting her friend Alice and told Fred that she would be visiting Alice. Fred, however, has forgotten what Harriet said. Without reflecting on it too much, Fred considers the world t possible, where Harriet said that she was visiting her brother Bob. Now, in fact, if Harriet had told Fred that she was visiting Bob, Fred would have remembered that fact, since Harriet had just had a fight with Bob the week before. Thus, in world t, Fred would not consider world s possible, since in world t, Fred would remember that Harriet said she was visiting Bob, rather than Alice. Perhaps with some introspection, Fred might realize that t is not possible, because in t he would have remembered what Harriet said. But people do not always do such introspection.

By investigating the properties of knowledge in a more general framework, as we do here, we can see how these properties depend on the assumptions we make about the possibility relations $\mathcal{K}_i$. In addition, we obtain general proof techniques, which in particular enable us to characterize in a precise sense the complexity of enable us to characterize in a precise sense the complexity of reasoning about knowledge.

This chapter is somewhat more technical than the previous ones; we have highlighted the major ideas in the text, and have left many of the details to the exercises. A reader interested just in the results may want to skip many of the proofs. However, we strongly encourage the reader who wants to gain a deeper appreciation of the techniques of modal logic to work through these exercises.

3.1 Completeness Results

As we said before, we begin by considering arbitrary Kripke structures, without the assumption that the possibility relations $\mathcal{K}_i$ are equivalence relations. Before we go on, we need to define some additional notation. Let $\mathcal{L}_n(\Phi)$ be the set of formulas that can be built up starting from the primitive propositions in Φ, using conjunction, negation, and the modal operators $K_1, \ldots, K_n$. Let $\mathcal{L}_n^D(\Phi)$ (resp., $\mathcal{L}_n^C(\Phi)$) be the language that results when we allow in addition the modal operators D_G (resp., operators E_G and C_G), where G is a nonempty subset of $\{1, \ldots, n\}$. In addition, we consider the language $\mathcal{L}_n^{CD}(\Phi)$, where formulas are formed using all the operators C_G, D_G, and E_G. Let $\mathcal{M}_n(\Phi)$ be the class of all Kripke structures for n agents over Φ (with no restrictions on the $\mathcal{K}_i$ relations). Later we consider various subclasses of $\mathcal{M}_n(\Phi)$, obtained by restricting the $\mathcal{K}_i$ relations appropriately. For example, we consider $\mathcal{M}_n^{rst}(\Phi)$, the Kripke structures where the $\mathcal{K}_i$ relation is reflexive, symmetric, and transitive (i.e., an equivalence relation); these are precisely the structures discussed in the previous chapter. For notational convenience, we take the set Φ of primitive propositions to be fixed from now on and suppress it from the notation, writing $\mathcal{L}_n$ instead of $\mathcal{L}_n(\Phi)$, $\mathcal{M}_n$ instead of $\mathcal{M}_n(\Phi)$, and so on.

If A is a set, define $|A|$ to be the cardinality of A (i.e., the number of elements in A). We define $|\varphi|$, the *length* of a formula $\varphi \in \mathcal{L}_n^{CD}$, to be the number of symbols that occur in φ; for example, $|p \wedge E_{\{1,2\}} p| = 9$. In general, the length of a formula of the form $C_G\psi$, $E_G\psi$, or $D_G\psi$ is $2 + 2|G| + |\psi|$, since we count the elements in G as distinct symbols, as well as the commas and set braces in G. We also define what it means for ψ to be a *subformula* of φ. Informally, ψ is a subformula of φ if it is a formula that is a substring of φ. The formal definition proceeds by induction on the structure of φ: ψ is a subformula of $\varphi \in \mathcal{L}_n$ if either (a) $\psi = \varphi$ (so that φ and ψ are syntactically identical), (b) φ is of the form $\neg\varphi'$, $K_i\varphi'$, $C_G\varphi'$, $D_G\varphi'$, or $E_G\varphi'$, and ψ is a subformula of φ', or (c) φ is of the form $\varphi' \wedge \varphi''$ and ψ is a subformula of either φ' or φ''. Let $Sub(\varphi)$ be the set of all subformulas of φ. We leave it to the reader to check that $|Sub(\varphi)| \leq |\varphi|$; that is, the length of φ is an upper bound on the number of subformulas of φ (Exercise 3.1).

Although we have now dropped the restriction that the $\mathcal{K}_i$'s be equivalence relations, the definition of what it means for a formula φ in $\mathcal{L}_n^{CD}$ (or any of its sublanguages) to be true at a state s in the Kripke structure $M \in \mathcal{M}_n$ remains the same, as do the notions of validity and satisfiability. Thus, for example, $(M, s) \models K_i\varphi$ (i.e., agent i knows φ at state s in M) exactly if φ is true at all the states t such that

$(s, t) \in \mathcal{K}_i$. We say that φ is *valid with respect to* $\mathcal{M}_n$, and write $\mathcal{M}_n \models \varphi$, if φ is valid in all the structures in $\mathcal{M}_n$. More generally, if $\mathcal{M}$ is some subclass of $\mathcal{M}_n$, we say that φ is *valid with respect to* $\mathcal{M}$, and write $\mathcal{M} \models \varphi$, if φ is valid in all the structures in $\mathcal{M}$. Similarly, we say that φ is satisfiable with respect to $\mathcal{M}$ if φ is satisfied in some structure in $\mathcal{M}$.

We are interested in characterizing the properties of knowledge in Kripke structures in terms of the formulas that are valid in Kripke structures. Note that we should expect *fewer* formulas to be valid than were valid in the Kripke structures considered in the previous chapter, for we have now dropped the restriction that the $\mathcal{K}_i$'s are equivalence relations. The class $\mathcal{M}_n^{rst}$ of structures is a proper subclass of $\mathcal{M}_n$. Therefore, a formula that is valid with respect to $\mathcal{M}_n$ is certainly valid with respect to the more restricted class $\mathcal{M}_n^{rst}$. As we shall see, the converse does not hold.

We start by considering the language $\mathcal{L}_n$; we deal with common knowledge and distributed knowledge later on. We observed in the previous chapter that the Distribution Axiom and the Knowledge Generalization Rule hold no matter how we modify the $\mathcal{K}_i$ relations. Thus, the following theorem should not come as a great surprise.

Theorem 3.1.1 *For all formulas $\varphi, \psi \in \mathcal{L}_n$, structures $M \in \mathcal{M}_n$, and agents $i = 1, \ldots, n$,*

(a) if φ is an instance of a propositional tautology, then $\mathcal{M}_n \models \varphi$,

(b) if $M \models \varphi$ and $M \models \varphi \Rightarrow \psi$ then $M \models \psi$,

(c) $\mathcal{M}_n \models (K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi$,

(d) if $M \models \varphi$ then $M \models K_i\varphi$.

Proof Parts (a) and (b) follow immediately from the fact that the interpretation of $\wedge$ and $\neg$ in the definition of $\models$ is the same as in propositional logic. The proofs of part (c) and (d) are identical to the proofs of parts (a) and (b) of Theorem 2.4.1. ∎

We now show that, in a precise sense, these properties completely characterize the formulas of $\mathcal{L}_n$ that are valid with respect to $\mathcal{M}_n$. To do so, we have to consider the notion of *provability*. An *axiom system* AX consists of a collection of *axioms* and *inference rules*. An axiom is a formula, and an inference rule has the form "from $\varphi_1, \ldots, \varphi_k$ infer ψ," where $\varphi_1, \ldots, \varphi_k, \psi$ are formulas. We are actually interested in (substitution) instances of axioms and inference rules (so we are really thinking

of axioms and inference rules as *schemes*). For example, the formula $K_1q \vee \neg K_1q$ is an instance of the propositional tautology $p \vee \neg p$, obtained by substituting K_1q for p. A *proof* in AX consists of a sequence of formulas, each of which is either an instance of an axiom in AX or follows by an application of an inference rule. (If "from $\varphi_1, \ldots, \varphi_k$ infer ψ" is an instance of an inference rule, and if the formulas $\varphi_1, \ldots, \varphi_k$ have appeared earlier in the proof, then we say that ψ *follows by an application of an inference rule*.) A proof is said to be a *proof of the formula* φ if the last formula in the proof is φ. We say φ is *provable in* AX, and write $AX \vdash \varphi$, if there is a proof of φ in AX.

Consider the following axiom system K_n, which consists of the two axioms and two inference rules given below:

A1. All tautologies of propositional calculus

A2. $(K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi, i = 1, \ldots, n$ (Distribution Axiom)

R1. From φ and $\varphi \Rightarrow \psi$ infer ψ (modus ponens)

R2. From φ infer $K_i\varphi, i = 1, \ldots, n$ (Knowledge Generalization)

Recall that we are actually interested in instances of axioms and inference rules. For example,

$$(K_1(p \wedge q) \wedge K_1((p \wedge q) \Rightarrow \neg K_2r)) \Rightarrow K_1\neg K_2r$$

is a substitution instance of the Distribution Axiom.

As a typical example of the use of K_n, consider the following proof of the formula $K_i(p \wedge q) \Rightarrow K_ip$. We give the axiom used or the inference rule applied and the lines it was applied to in parentheses at the end of each step:

1. $(p \wedge q) \Rightarrow p$ (A1)

2. $K_i((p \wedge q) \Rightarrow p)$ (1,R2)

3. $(K_i(p \wedge q) \wedge K_i((p \wedge q) \Rightarrow p)) \Rightarrow K_ip$ (A2)

4. $((K_i(p \wedge q) \wedge K_i((p \wedge q) \Rightarrow p)) \Rightarrow K_ip)$
   $\Rightarrow (K_i((p \wedge q) \Rightarrow p) \Rightarrow (K_i(p \wedge q) \Rightarrow K_ip))$
   (A1, since this is an instance of the propositional tautology
   $((p_1 \wedge p_2) \Rightarrow p_3) \Rightarrow (p_2 \Rightarrow (p_1 \Rightarrow p_3)))$

5. $K_i((p \wedge q) \Rightarrow p) \Rightarrow (K_i(p \wedge q) \Rightarrow K_i p)$ (3,4,R1)

6. $K_i(p \wedge q) \Rightarrow K_i p$ (2,5,R1)

This proof already shows how tedious the proof of even simple formulas can be. Typically we tend to combine several steps when writing up a proof, especially those that involve only propositional reasoning (A1 and R1).

The reader familiar with formal proofs in propositional or first-order logic should be warned that one technique that works in these cases, namely, the use of the *deduction theorem*, does *not* work for K_n. To explain the deduction theorem, we need one more definition. We generalize the notion of provability by defining φ to be *provable from* ψ *in the axiom system* AX, written $AX, \psi \vdash \varphi$, if there is a sequence of steps ending with φ, each of which is either an instance of an axiom of AX, ψ itself, or follows from previous steps by an application of an inference rule of AX. The deduction theorem is said to hold for AX if $AX, \psi \vdash \varphi$ implies $AX \vdash \psi \Rightarrow \varphi$. Although the deduction theorem holds for the standard axiomatizations of propositional logic and first-order logic, it does not hold for K_n. To see this, observe that for any formula φ, by an easy application of Knowledge Generalization (R2) we have $\mathrm{K}_n, \varphi \vdash K_i\varphi$. However, we do not in general have $\mathrm{K}_n \vdash \varphi \Rightarrow K_i\varphi$: it is certainly not the case in general that if φ is true, then agent i knows φ. It turns out that the Knowledge Generalization Rule is essentially the cause of the failure of the deduction theorem for K_n. This issue is discussed in greater detail in Exercises 3.8 and 3.29.

We return now to our main goal, that of proving that K_n characterizes the set of formulas that are valid with respect to $\mathcal{M}_n$. An axiom system AX is said to be *sound* for a language $\mathcal{L}$ with respect to a class $\mathcal{M}$ of structures if every formula in $\mathcal{L}$ provable in AX is valid with respect to $\mathcal{M}$. The system AX is *complete* for $\mathcal{L}$ with respect to $\mathcal{M}$ if every formula in $\mathcal{L}$ that is valid with respect to $\mathcal{M}$ is provable in AX. We think of AX as characterizing the class $\mathcal{M}$ if it provides a sound and complete axiomatization of that class; notationally, this amounts to saying that for all formulas φ, we have $AX \vdash \varphi$ if and only if $\mathcal{M} \models \varphi$. Soundness and completeness provide a tight connection between the *syntactic* notion of provability and the *semantic* notion of validity.

We plan to show that K_n provides a sound and complete axiomatization for $\mathcal{L}_n$ with respect to $\mathcal{M}_n$. We need one more round of definitions in order to do this. Given an axiom system AX, we say a formula φ is AX-*consistent* if $\neg\varphi$ is not provable in AX. A finite set $\{\varphi_1, \ldots, \varphi_k\}$ of formulas is AX-consistent exactly if the

conjuction $\varphi_1 \wedge \ldots \wedge \varphi_k$ of its members is AX-consistent. As is standard, we take the empty conjunction to be the formula *true,* so the empty set is AX-consistent exactly if *true* is AX-consistent. An infinite set of formulas is AX-consistent exactly if all of its finite subsets are AX-consistent. Recall that a language is a set of formulas. A set F of formulas is a *maximal AX-consistent set* with respect to a language $\mathcal{L}$ if (1) it is AX-consistent, and (2) for all φ in $\mathcal{L}$ but not in F, the set $F \cup \{\varphi\}$ is not AX-consistent.

Lemma 3.1.2 *Suppose that the language $\mathcal{L}$ consists of a countable set of formulas and is closed with respect to propositional connectives (so that if φ and ψ are in the language, then so are $\varphi \wedge \psi$ and $\neg\varphi$). In a consistent axiom system AX that includes every instance of A1 and R1 for the language $\mathcal{L}$, every AX-consistent set $F \subseteq \mathcal{L}$ can be extended to a maximal AX-consistent set with respect to $\mathcal{L}$. In addition, if F is a maximal AX-consistent set, then it satisfies the following properties:*

(a) for every formula $\varphi \in \mathcal{L}$, exactly one of φ and $\neg\varphi$ is in F,

(b) $\varphi \wedge \psi \in F$ iff $\varphi \in F$ and $\psi \in F$,

(c) if φ and $\varphi \Rightarrow \psi$ are both in F, then ψ is in F,

(d) if φ is provable in AX, then $\varphi \in F$.

Proof Let F be an AX-consistent subset of formulas in $\mathcal{L}$. To show that F can be extended to a maximal AX-consistent set, we first construct a sequence $F_0, F_1, F_2, \ldots$ of AX-consistent sets as follows. Because $\mathcal{L}$ is a countable language, let $\psi_1, \psi_2, \ldots$ be an enumeration of the formulas in $\mathcal{L}$. Let $F_0 = F$, and inductively construct the rest of the sequence by taking $F_{i+1} = F_i \cup \{\psi_{i+1}\}$ if this set is AX-consistent and otherwise by taking $F_{i+1} = F_i$. It is easy to see that each set in the sequence $F_0, F_1, \ldots$ is AX-consistent, and that this is a nondecreasing sequence of sets. Let $F = \cup_{i=0}^{\infty} F_i$. Each finite subset of F must be contained in F_j for some j, and thus must be AX-consistent (since F_j is AX-consistent). It follows that F itself is AX-consistent. We claim that in fact F is a maximal AX-consistent set. For suppose $\psi \in \mathcal{L}$ and $\psi \notin F$. Since ψ is a formula in $\mathcal{L}$, it must appear in our enumeration, say as ψ_k. If $F_k \cup \{\psi_k\}$ were AX-consistent, then our construction would guarantee that $\psi_k \in F_{k+1}$, and hence that $\psi_k \in F$. Because $\psi_k = \psi \notin F$, it follows that $F_k \cup \{\psi\}$ is not AX-consistent. Hence $F \cup \{\psi\}$ is also not AX-consistent. It follows that F is a maximal AX-consistent set.

To see that maximal AX-consistent sets have all the properties we claimed, let F be a maximal AX-consistent set. If $\varphi \in \mathcal{L}$, we now show that one of $F \cup \{\varphi\}$ and $F \cup \{\neg\varphi\}$ is AX-consistent. For assume to the contrary that neither of them is AX-consistent. It is not hard to see that $F \cup \{\varphi \vee \neg\varphi\}$ is then also not AX-consistent (Exercise 3.2). So F is not AX-consistent, because $\varphi \vee \neg\varphi$ is a propositional tautology. This gives a contradiction. If $F \cup \{\varphi\}$ is AX-consistent, then we must have $\varphi \in F$ since F is a maximal AX-consistent set. Similarly, if $F \cup \{\neg\varphi\}$ is AX-consistent then $\neg\varphi \in F$. Thus, one of φ or $\neg\varphi$ is in F. It is clear that we cannot have both φ and $\neg\varphi$ in F, for otherwise F would not be AX-consistent. This proves (a).

Part (a) is enough to let us prove all the other properties we claimed. For example, if $\varphi \wedge \psi \in F$, then we must have $\varphi \in F$, for otherwise, as we just showed, we would have $\neg\varphi \in F$, and F would not be AX-consistent. Similarly, we must have $\psi \in F$. Conversely, if φ and ψ are both in F, we must have $\varphi \wedge \psi \in F$, for otherwise we would have $\neg(\varphi \wedge \psi) \in F$, and, again, F would not be AX-consistent. We leave the proof that F has properties (c) and (d) to the reader (Exercise 3.3). ∎

We can now prove that K_n is sound and complete.

Theorem 3.1.3 K_n *is a sound and complete axiomatization with respect to* $\mathcal{M}_n$ *for formulas in the language* $\mathcal{L}_n$.

Proof Using Theorem 3.1.1, it is straightforward to prove by induction on the length of a proof of φ that if φ is provable in K_n, then φ is valid with respect to $\mathcal{M}_n$ (see Exercise 3.4). It follows that K_n is sound with respect to $\mathcal{M}_n$.

To prove completeness, we must show that every formula in $\mathcal{L}_n$ that is valid with respect to $\mathcal{M}_n$ is provable in K_n. It suffices to prove that

Every K_n-consistent formula in $\mathcal{L}_n$ is satisfiable with respect to $\mathcal{M}_n$. $\quad (*)$

For suppose we can prove $(*)$, and φ is a valid formula in $\mathcal{L}_n$. If φ is not provable in K_n, then neither is $\neg\neg\varphi$, so, by definition, $\neg\varphi$ is K_n-consistent. It follows from $(*)$ that $\neg\varphi$ is satisfiable with respect to $\mathcal{M}_n$, contradicting the validity of φ with respect to $\mathcal{M}_n$.

We prove $(*)$ using a general technique that works for a wide variety of modal logics. We construct a special structure $M^c \in \mathcal{M}_n$, called the *canonical structure for* K_n. M^c has a state s_V corresponding to every maximal K_n-consistent set V. Then we show

$$(M^c, s_V) \models \varphi \text{ iff } \varphi \in V. \qquad (**)$$

That is, we show that a formula is true at a state s_V exactly if it is one of the formulas in V. Note that $(**)$ suffices to prove $(*)$, for by Lemma 3.1.2, if φ is K_n-consistent, then φ is contained in some maximal K_n-consistent set V. From $(**)$ it follows that $(M^c, s_V) \models \varphi$, and so φ is satisfiable in M^c. Therefore, φ is satisfiable with respect to $\mathcal{M}_n$, as desired.

We proceed as follows. Given a set V of formulas, define $V/K_i = \{\varphi \mid K_i\varphi \in V\}$. For example, if $V = \{K_1 p, K_2K_1q, K_1K_3p \wedge q, K_1K_3q\}$, then $V/K_1 = \{p, K_3q\}$. Let $M^c = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where

$$\begin{aligned} S &= \{s_V \mid V \text{ is a maximal } \mathrm{K}_n\text{-consistent set}\} \\ \pi(s_V)(p) &= \begin{cases} \textbf{true} & \text{if } p \in V \\ \textbf{false} & \text{if } p \notin V \end{cases} \\ \mathcal{K}_i &= \{(s_V, s_W) \mid V/K_i \subseteq W\}. \end{aligned}$$

We now show that for all $s_v \in S$ we have $(M^c, s_V) \models \varphi$ iff $\varphi \in V$. We proceed by induction on the structure of formulas. More precisely, assuming that the claim holds for all subformulas of φ, we also show that it holds for φ.

If φ is a primitive proposition p, this is immediate from the definition of $\pi(s_V)$. The cases where φ is a conjunction or a negation are simple and left to the reader (Exercise 3.5). Assume that φ is of the form $K_i\psi$ and that $\varphi \in V$. Then $\psi \in V/K_i$ and, by definition of $\mathcal{K}_i$, if $(s_V, s_W) \in \mathcal{K}_i$, then $\psi \in W$. Thus, using the induction hypothesis, $(M^c, s_W) \models \psi$ for all W such that $(s_V, s_W) \in \mathcal{K}_i$. By the definition of $\models$, it follows that $(M^c, s_V) \models K_i\psi$.

For the other direction, assume $(M^c, s_V) \models K_i\psi$. It follows that the set $(V/K_i) \cup \{\neg\psi\}$ is not K_n-consistent. For suppose otherwise. Then, by Lemma 3.1.2, it would have a maximal K_n-consistent extension W and, by construction, we would have $(s_V, s_W) \in \mathcal{K}_i$. By the induction hypothesis we have $(M^c, s_W) \models \neg\psi$, and so $(M^c, s_V) \models \neg K_i\psi$, contradicting our original assumption. Since $(V/K_i) \cup \{\neg\psi\}$ is not K_n-consistent, there must be some finite subset, say $\{\varphi_1, \ldots, \varphi_k, \neg\psi\}$, which is not K_n-consistent. Thus, by propositional reasoning (Exercise 3.6), we have

$$\mathrm{K}_n \vdash \varphi_1 \Rightarrow (\varphi_2 \Rightarrow (\ldots \Rightarrow (\varphi_k \Rightarrow \psi)\ldots)).$$

By R2, we have

$$\mathrm{K}_n \vdash K_i(\varphi_1 \Rightarrow (\varphi_2 \Rightarrow (\ldots \Rightarrow (\varphi_k \Rightarrow \psi)\ldots))).$$

By induction on k, together with axiom A2 and propositional reasoning, we can show (Exercise 3.7)

$$\begin{aligned}\mathrm{K}_n \vdash K_i(\varphi_1 \Rightarrow (\varphi_2 \Rightarrow (\ldots \Rightarrow (\varphi_k \Rightarrow \psi)\ldots)) \\ \Rightarrow (K_i\varphi_1 \Rightarrow (K_i\varphi_2 \Rightarrow (\ldots \Rightarrow (K_i\varphi_k \Rightarrow K_i\psi)\ldots)).\end{aligned}$$

Now from R1, we get

$$\mathrm{K}_n \vdash K_i\varphi_1 \Rightarrow (K_i\varphi_2 \Rightarrow (\ldots \Rightarrow (K_i\varphi_k \Rightarrow K_i\psi)\ldots)).$$

By part (d) of Lemma 3.1.2, it follows that

$$K_i\varphi_1 \Rightarrow (K_i\varphi_2 \Rightarrow (\ldots \Rightarrow (K_i\varphi_k \Rightarrow K_i\psi)\ldots)) \in V.$$

Because $\varphi_1, \ldots, \varphi_k \in V/K_i$, we must have $K_i\varphi_1, \ldots, K_i\varphi_k \in V$. By part (c) of Lemma 3.1.2, applied repeatedly, it follows that $K_i\psi \in V$, as desired. ∎

We have thus shown that K_n completely characterizes the formulas in $\mathcal{L}_n$ that are valid with respect to $\mathcal{M}_n$, where there are no restrictions on the $\mathcal{K}_i$ relations. What happens if we restrict the $\mathcal{K}_i$ relations? In Chapter 2, we observed that we do get extra properties if we take the $\mathcal{K}_i$ relations to be reflexive, symmetric, and transitive. These properties are the following:

A3. $K_i\varphi \Rightarrow \varphi, \quad i = 1, \ldots, n$ (Knowledge Axiom)

A4. $K_i\varphi \Rightarrow K_iK_i\varphi, \quad i = 1, \ldots, n$ (Positive Introspection Axiom)

A5. $\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi, \quad i = 1, \ldots, n$ (Negative Introspection Axiom)

We remarked earlier that axiom A3 has been taken by philosophers to capture the difference between knowledge and belief. From this point of view, the man we spoke of at the beginning of the chapter who "knew" his son was drug-free should really be said to *believe* his son was drug-free, but not to know it. If we want to model such a notion of belief, then (at least according to some philosophers) we ought to drop A3, but add an axiom that says that an agent does not believe *false*:

A6. $\neg K_i(\mathit{false}), \quad i = 1, \ldots, n$ (Consistency Axiom)

It is easy to see that A6 is provable from A3, A1, and R1 (see Exercise 3.9).

Historically, axiom A2 has been called **K**, A3 has been called **T**, A4 has been called **4**, A5 has been called **5**, and A6 has been called **D**. We get different modal logics by considering various subsets of these axioms. These logics have typically been named after the significant axioms they use. For example, in the case of one agent, the system with axioms and rules A1, A2, R1, and R2 has been called K, since its most significant axiom is **K**. Similarly, the axiom system KD45 is the result of combining the axioms **K**, **D**, **4**, and **5** with A1, R1, and R2, and KT4 is the result of combining the axioms **K**, **T**, and **4** with A1, R1, and R2. Some of the axiom systems are commonly called by other names as well. The K is quite often omitted, so that KT becomes T, KD becomes D, and so on; KT4 has traditionally been called S4 and KT45 has been called S5. (The axioms **K**, **T**, **4**, and **5**, together with rule R2, are what we called the S5 properties in Chapter 2.) We stick with the traditional names here for those logics that have them, since they are in common usage, except that we use the subscript n to emphasize the fact that we are considering systems with n agents rather than only one agent. Thus, for example, we speak of the logics T_n or $\mathrm{S5}_n$. We occasionally omit the subscript if $n = 1$, in line with more traditional notation.

Philosophers have spent years arguing which of these axioms, if any, best captures the knowledge of an agent. We do not believe that there is one "true" notion of knowledge; rather, the appropriate notion depends on the application. As we said in Chapter 2, for many of our applications the axioms of S5 seem most appropriate (although philosophers have argued quite vociferously against them, particularly axiom A5). Rather than justify these axioms further, we focus here on the relationship between these axioms and the properties of the $\mathcal{K}_i$ relation, and on the effect of this relationship on the difficulty of reasoning about knowledge. (Some references on the issue of justification of the axioms can be found in the bibliographic notes at the end of the chapter.) Since we do not have the space to do an exhaustive study of all the logics that can be formed by considering all possible subsets of the axioms, we focus on some representative cases here, namely K_n, T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, and $\mathrm{KD45}_n$. These provide a sample of the logics that have been considered in the literature and demonstrate some of the flexibility of this general approach to modeling knowledge. K_n is the minimal system, and it enables us to study what happens when there are in some sense as few restrictions as possible on the K_i operator, given our possible-worlds framework. The minimal extension of K_n that requires that what is known is necessarily true is the system T_n. Researchers who have accepted the arguments against A5 but have

otherwise been happy with the axioms of $S5_n$ have tended to focus on $S4_n$. On the other hand, researchers who were willing to accept the introspective properties embodied by A4 and A5, but wanted to consider belief rather than knowledge, have tended to consider KD45 or K45. For definiteness, we focus on KD45 here, but all our results for KD45 carry over with very little change to K45.

Theorem 3.1.3 implies that the formulas provable in K_n are precisely those that are valid with respect to $\mathcal{M}_n$. We want to connect the remaining axioms with various restrictions on the possibility relations $\mathcal{K}_i$. We have already considered one possible restriction on the $\mathcal{K}_i$ relations (namely, that they be reflexive, symmetric, and transitive). We now consider others. We say that a binary relation $\mathcal{K}$ on a set S is *Euclidean* if, for all $s, t, u \in S$, whenever $(s, t) \in \mathcal{K}$ and $(s, u) \in \mathcal{K}$, then $(t, u) \in \mathcal{K}$; we say that $\mathcal{K}$ is *serial* if, for all $s \in S$, there is some t such that $(s, t) \in \mathcal{K}$.

Some of the relationships between various conditions we can place on binary relations are captured in the following lemma, whose proof is left to the reader (Exercise 3.12).

Lemma 3.1.4

(a) If $\mathcal{K}$ is reflexive and Euclidean, then $\mathcal{K}$ is symmetric and transitive.

(b) If $\mathcal{K}$ is symmetric and transitive, then $\mathcal{K}$ is Euclidean.

(c) The following are equivalent:

(i) $\mathcal{K}$ is reflexive, symmetric, and transitive.

(ii) $\mathcal{K}$ is symmetric, transitive, and serial.

(iii) $\mathcal{K}$ is reflexive and Euclidean.

Let $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$; $\mathcal{M}_n^{rst}$; $\mathcal{M}_n^{elt}$) be the class of all structures for n agents where the possibility relations are reflexive (resp., reflexive and transitive; reflexive, symmetric, and transitive; Euclidean, serial, and transitive). As we observed earlier, since an equivalence relation is one that is reflexive, symmetric, and transitive, $\mathcal{M}_n^{rst}$ is precisely the class of structures we considered in Chapter 2.

The next theorem shows a close connection between various combinations of axioms, on the one hand, and various restrictions on the possibility relations $\mathcal{K}_i$, on the other hand. For example, axiom A3 (the Knowledge Axiom $K_i\varphi \Rightarrow \varphi$) corresponds to reflexivity of $\mathcal{K}_i$. To demonstrate one part of this correspondence,

we now show that axiom A3 is valid in all structures in $\mathcal{M}_n^r$. If s is a world in a structure $M \in \mathcal{M}_n^r$, then agent i must consider s to be one of his possible worlds in s. Thus, if agent i knows φ in s, then φ must be true in s; that is, $(M, s) \models K_i\varphi \Rightarrow \varphi$. Therefore, T_n is sound with respect to $\mathcal{M}_n^r$. We might hope that, conversely, every structure that satisfies all instances of axiom A3 is in $\mathcal{M}_n^r$. Unfortunately, this is not the case (we return to this point a little later). Nevertheless, as we shall see in the proof of the next theorem, axiom A3 forces the possibility relations in the canonical structure to be reflexive. As we shall see, this is sufficient to prove that T_n is complete with respect to $\mathcal{M}_n^r$.

Theorem 3.1.5 *For formulas in the language $\mathcal{L}_n$:*

(a) T_n *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^r$,

(b) $\mathrm{S4}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{rt}$,

(c) $\mathrm{S5}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{rst}$,

(d) $\mathrm{KD45}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{elt}$.

Proof We first consider part (a). We already showed that T_n is sound with respect to $\mathcal{M}_n^r$. For completeness, we need to show that every T_n-consistent formula is satisfiable in some structure in $\mathcal{M}_n^r$. This is done exactly as in the proof of Theorem 3.1.3. We define a canonical structure M^c for T_n each of whose states corresponds to a maximal T_n-consistent set V of formulas. The $\mathcal{K}_i$ relations are defined as in the proof of Theorem 3.1.3, namely, $(s_V, s_W) \in \mathcal{K}_i$ in M^c exactly if $V/K_i \subseteq W$, where $V/K_i = \{\varphi \mid K_i\varphi \in V\}$. A proof identical to that of Theorem 3.1.3 can now be used to show that $\varphi \in V$ iff $(M^c, s_V) \models \varphi$, for all maximal T_n-consistent sets V. Furthermore, it is easy to see that every maximal T_n-consistent set V contains every instance of axiom A3. Therefore, all instances of axiom A3 are true at s_V. It follows immediately that $V/K_i \subseteq V$. So by definition of $\mathcal{K}_i$, it follows that $(s_V, s_V) \in \mathcal{K}_i$. So $\mathcal{K}_i$ is indeed reflexive, and hence $M^c \in \mathcal{M}_n^r$. Assume now that φ is a T_n-consistent formula. As in the proof of Theorem 3.1.3, it follows that φ is satisfiable in M^c. Since, as we just showed, $M^c \in \mathcal{M}_n^r$, it follows that φ is satisfiable in some structure in $\mathcal{M}_n^r$, as desired. This completes the proof of part (a).

To prove part (b), we show that just as axiom A3 corresponds to reflexivity, similarly axiom A4 corresponds to transitivity. It is easy to see that A4 is valid in all structures where the possibility relation is transitive. Moreover, A4 forces the

possibility relations in the canonical structure to be transitive. To see this, suppose that $(s_V, s_W), (s_W, s_X) \in \mathcal{K}_i$ and that all instances of A4 are true at s_V. Then if $K_i\varphi \in V$, by A4 we have $K_iK_i\varphi \in V$, and, by the construction of M^c, we have $K_i\varphi \in W$ and $\varphi \in X$. Thus, $V/K_i \subseteq X$ and $(s_V, s_X) \in \mathcal{K}_i$, as desired. That means that in the canonical structure for $S4_n$, the possibility relation is both reflexive and transitive, so the canonical structure is in $\mathcal{M}_n^{rt}$. The proof is now very similar to that of part (a).

The proof of parts (c) and (d) go in the same way. Here the key correspondences are that axiom A5 corresponds to a Euclidean possibility relation and axiom A6 corresponds to a serial relation (Exercise 3.13). ∎

We say that a structure M is a *model of* K_n if every formula provable in K_n is valid in M. We can similarly say that a structure is a model of T_n, $S4_n$, $S5_n$, and $KD45_n$. The soundness part of Theorem 3.1.5 shows that every structure in $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$, $\mathcal{M}_n^{elt}$) is a model of T_n (resp., $S4_n$, $S5_n$, $KD45_n$). We might be tempted to conjecture that the converse also holds, so that, for example, if a structure is a model of $S5_n$, then it is in $\mathcal{M}_n^{rst}$. This is not quite true, as the following example shows. Suppose that $n = 1$ and $\Phi = \{p\}$, and let M be the structure consisting of two states s and t, such that $\pi(s)(p) = \pi(t)(p) = \mathbf{true}$ and $\mathcal{K}_1 = \{(s, t), (t, t)\}$, as shown in Figure 3.1.

Figure 3.1 A model of $S5_1$ that is not in $\mathcal{M}_1^{rst}$

The structure M is not in $\mathcal{M}_1^r$, let alone $\mathcal{M}_1^{rst}$, but it is easy to see that it is a model of $S5_1$ and *a fortiori* a model of $S4_1$ and T_1 (Exercise 3.15). Nevertheless, the intuition behind the conjecture is almost correct. In fact, it is correct in two senses. If s is a state in a Kripke structure M, and s' is a state in a Kripke structure M', then we say that (M, s) and (M', s') are *equivalent*, and write $(M, s) \equiv (M', s')$, if they satisfy exactly the same formulas in the language $\mathcal{L}_n$. That is, $(M, s) \equiv (M', s')$ if, for all formulas $\varphi \in \mathcal{L}_n$, we have $(M, s) \models \varphi$ if and only if $(M', s') \models \varphi$. One sense in which the previous conjecture is correct is that every model M of T_n (resp., $S4_n$, $S5_n$, $KD45_n$) can effectively be converted to a structure M' in $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$,

$\mathcal{M}_n^{elt}$) with the same state space, such that $(M, s) \equiv (M', s)$ for every state s (see Exercise 3.16).

The second sense in which the conjecture is correct involves the notion of a *frame*. We define a *frame for n agents* to be a tuple $(S, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where S is a set of states and $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are binary relations on S. Thus, a frame is like a Kripke structure without the function π. Notice that the Aumann structures defined in Section 2.5 can be viewed as frames. We say that the Kripke structure $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ is *based on* the frame $(S, \mathcal{K}_1, \ldots, \mathcal{K}_n)$. A formula φ is *valid* in frame F if it is valid in every Kripke structure based on F. It turns out that if we look at the level of frames rather than at the level of structures, then we get what can be viewed as a partial converse to Theorem 3.1.5. For example, the $\mathcal{K}_i$'s in a frame F are reflexive *if and only if* every instance of the Knowledge Axiom A3 is valid in F. This suggests that the axioms are tied more closely to frames than they are to structures. Although we have shown that, for example, we can find a structure that is a model of $S5_n$ but is not in $\mathcal{M}_n^{rst}$ (or even $\mathcal{M}_n^r$), this is not the case at the level of frames. If a frame is a model of $S5_n$, then it must be in $\mathcal{F}_n^{rst}$. Conversely, if a frame is in $\mathcal{F}_n^{rst}$, then it is a model of $S5_n$. See Exercise 3.17 for more details.

The previous results show the connection between various restrictions on the $\mathcal{K}_i$ relations and properties of knowledge. In particular, we have shown that A3 corresponds to reflexive possibility relations, A4 to transitive possibility relations, A5 to Euclidean possibility relations, and A6 to serial possibility relations.

Up to now we have not considered symmetric relations. It is not hard to check (using arguments similar to those used previously) that symmetry of the possibility relations corresponds to the following axiom:

A7. $\varphi \Rightarrow K_i \neg K_i \neg \varphi, \quad i = 1, \ldots, n$

Axiom A7 can also easily be shown to be a consequence of A3 and A5, together with propositional reasoning (Exercise 3.18). This corresponds to the observation made in Lemma 3.1.4 that a reflexive Euclidean relation is also symmetric. Since a reflexive Euclidean relation is also transitive, the reader may suspect that A4 is redundant in the presence of A3 and A5. This is essentially true. It can be shown that A4 is a consequence of A1, A2, A3, A5, R1, and R2 (see Exercise 3.19). Thus we can obtain an axiom system equivalent to $S5_n$ by eliminating A4; indeed, by using the observations of Lemma 3.1.4, we can obtain a number of axiomatizations that are equivalent to $S5_n$ (Exercise 3.20).

The preceding discussion is summarized by Table 3.1, which describes the correspondence between the axioms and the properties of the $\mathcal{K}_i$ relations.

Axiom	Property of $\mathcal{K}_i$
A3. $K_i\varphi \Rightarrow \varphi$	reflexive
A4. $K_i\varphi \Rightarrow K_iK_i\varphi$	transitive
A5. $\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$	Euclidean
A6. $\neg K_i\mathit{false}$	serial
A7. $\varphi \Rightarrow K_i\neg K_i\neg\varphi$	symmetric

Table 3.1 The correspondence between axioms and properties of $\mathcal{K}_i$

We conclude this section by taking a closer look at the single-agent case of S5 and KD45. The following result shows that in the case of S5 we can further restrict our attention to structures where the possibility relation is *universal*; that is, in every state, all states are considered possible. Intuitively, this means that in the case of S5 we can talk about *the* set of worlds the agent considers possible; this set is the same in every state and consists of all the worlds. Similarly, for KD45 we can restrict attention to structures with one distinguished state, which intuitively is the "real" world, and a set of states (which does not in general include the real world) corresponding to the worlds that the agent thinks possible in every state.

Proposition 3.1.6

(a) Assume that $M \in \mathcal{M}_1^{rst}$ and s is a state of M. Then there is a structure $M' = (S', \pi', \mathcal{K}_1')$, where $\mathcal{K}_1'$ is universal, that is, $\mathcal{K}_1' = \{(s,t) \mid s,t \in S'\}$, and a state s' of M' such that $(M,s) \equiv (M',s')$.

(b) Assume that $M \in \mathcal{M}_1^{elt}$ and s_0 is a state of M. Then there is a structure $M' = (\{s_0\} \cup S', \pi', \mathcal{K}_1')$, where S' is nonempty and $\mathcal{K}_1' = \{(s,t) \mid s \in \{s_0\} \cup S'$ and $t \in S'\}$, and a state s' of M' such that $(M,s_0) \equiv (M',s')$.

Proof We first consider part (b). Assume that $M = (S, \pi, \mathcal{K}_1) \in \mathcal{M}_1^{elt}$ and that $s_0 \in S$. Let $\mathcal{K}_1(s_0) = \{t \mid (s_0,t) \in \mathcal{K}_1\}$. Since $\mathcal{K}_1$ is serial, $\mathcal{K}_1(s_0)$ must be nonempty. It is also easy to check that since $\mathcal{K}_1$ is Euclidean, we have $(s,t) \in \mathcal{K}_1$ for all $s,t \in \mathcal{K}_1(s_0)$. Finally, since $\mathcal{K}_1$ is transitive, if $s \in \mathcal{K}_1(s_0)$ and $(s,t) \in \mathcal{K}_1$, then

$t \in \mathcal{K}_1(s_0)$. Let $M' = (\{s_0\} \cup \mathcal{K}_1(s_0), \pi', \mathcal{K}_1')$, where π' is the restriction of π to $\{s_0\} \cup \mathcal{K}_1(s_0)$, and $\mathcal{K}_1' = \{(s,t) \mid s \in \{s_0\} \cup \mathcal{K}_1(s_0) \text{ and } t \in \mathcal{K}_1(s_0)\}$. By the previous observations, $\mathcal{K}_1'$ is the restriction of $\mathcal{K}_1$ to $\{s_0\} \cup \mathcal{K}_1(s_0)$. Note that $\mathcal{K}_1'$ is serial (because $\mathcal{K}_1(s_0)$ is nonempty), Euclidean, and transitive. A straightforward induction on the structure of formulas now shows that for all $s \in \{s_0\} \cup \mathcal{K}_1(s_0)$ and all formulas $\varphi \in \mathcal{L}_n$, we have $(M, s) \models \varphi$ iff $(M', s) \models \varphi$. We leave details to the reader (Exercise 3.21).

For part (a), we proceed in the same way, except that we start with a structure $M \in \mathcal{M}_1^{rst}$. Using the fact that $\mathcal{K}_1$ is now reflexive, it is easy to show that the relation $\mathcal{K}_1'$ we construct is universal. The rest of the proof proceeds as before. ∎

It follows from Proposition 3.1.6 that we can assume without loss of generality that models of S5 have a particularly simple form, namely (S, π), where we do not mention the $\mathcal{K}_1$ relation but simply assume that $(s,t) \in \mathcal{K}_1$ for all $s, t \in S$. Similarly, we can take models of KD45 to have the form (s_0, S, π), where, as already discussed, the intuition is that s_0 is the "real" world, and S is the set of worlds that the agent considers possible. As we shall see, this simple representation of models for S5 and KD45 has important implications when it comes to the difficulty of deciding whether a formula is provable in S5 or KD45.

There is a similar simple representation for models of K45 (Exercise 3.22). We cannot in general get such simple representations for the other logics we have considered, nor can we get them even for $S5_n$ or $KD45_n$ if $n > 1$, that is, if we have two or more agents in the picture. For more information on the single-agent case of S5, see Exercise 3.23.

3.2 Decidability

In the preceding section we showed that the set of valid formulas of $\mathcal{M}_n$ is indeed characterized by K_n, and that the valid formulas of various interesting subclasses of $\mathcal{M}_n$ are characterized by other systems, such as T_n, $S4_n$, and $S5_n$. Our results, however, were not constructive; they gave no indication of how to tell whether a given formula was indeed provable (and thus also valid in the appropriate class of structures).

In this section, we present results showing that the question of whether a formula is valid is *decidable*; that is, there is an algorithm that, given as input a formula φ, will decide whether φ is valid. (It is beyond the scope of this book to give a formal