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

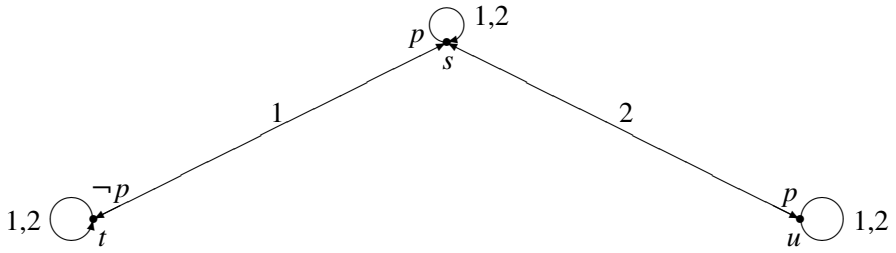

Figure 2.1 A simple Kripke structure

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).

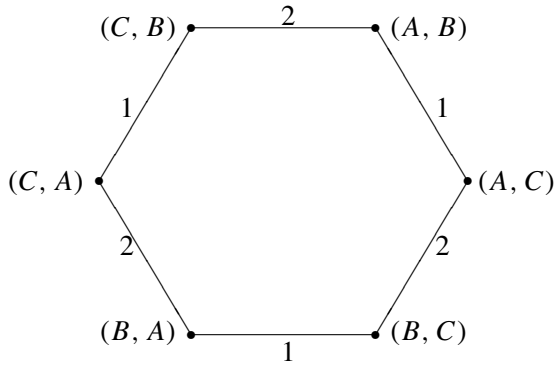

Figure 2.2 The Kripke structure describing a simple card game

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 $p_2 \Rightarrow K_1 p_2$,

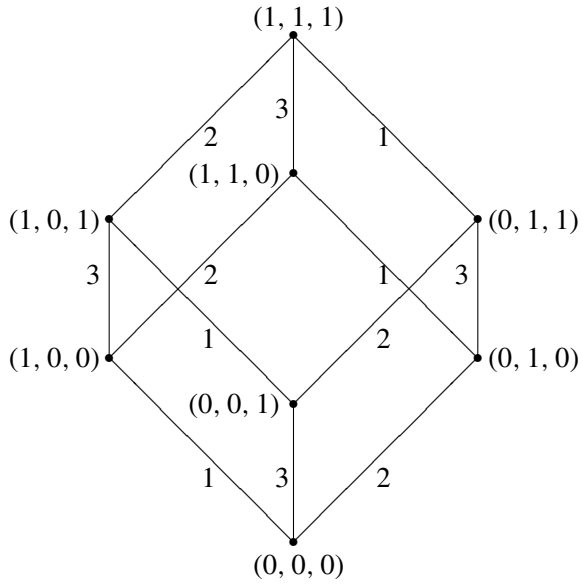

Figure 2.3 The Kripke structure for the muddy children puzzle with $n = 3$

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.

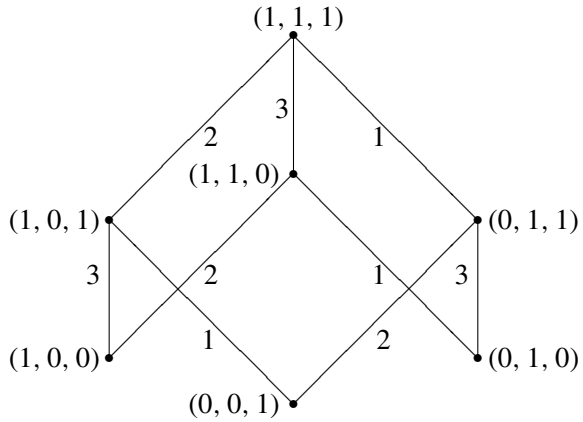

Figure 2.4 The Kripke structure after the father speaks

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.

2.5 An Event-Based Approach

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

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

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

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

S_j are called the *cells* of the partition $\mathcal{P}_i$, or the *information sets* of agent i. The intuition is that if S_j is an information set of agent i, and if $s \in S_j$, then the set of states that agent i considers possible (which corresponds to the information of agent i) is precisely S_j.

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

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

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

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

The event of *everyone in a group G knowing e* is captured by an operator $\mathsf{E}_G : 2^S \to 2^S$ defined as follows:

$$\mathsf{E}_G(e) = \bigcap_{i \in G} \mathsf{K}_i(e).$$

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

on ad infinitum. Formally, we define

$$\mathsf{C}_G(e) = \bigcap_{k=1}^{\infty} \mathsf{E}_G^k(e).$$

Finally, *distributed knowledge of an event e among the agents in a group G*, denoted $\mathsf{D}_G(e)$, is defined by

$$\mathsf{D}_G(e) = \{s \in S \mid (\bigcap_{i \in G} \mathcal{P}_i(s)) \subseteq e\}.$$

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

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

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

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

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

Proof See Exercise 2.15. ∎

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

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

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

- $\mathtt{ev}_M(p) = e_p^M$
- $\mathtt{ev}_M(\psi_1 \wedge \psi_2) = \mathtt{ev}_M(\psi_1) \cap \mathtt{ev}_M(\psi_2)$
- $\mathtt{ev}_M(\neg\psi) = S - \mathtt{ev}_M(\psi)$
- $\mathtt{ev}_M(K_i\psi) = \mathsf{K}_i(\mathtt{ev}_M(\psi))$
- $\mathtt{ev}_M(C_G\psi) = \mathsf{C}_G(\mathtt{ev}_M(\psi))$
- $\mathtt{ev}_M(D_G\psi) = \mathsf{D}_G(\mathtt{ev}_M(\psi))$

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

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

Proof See Exercise 2.17. ∎

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

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

Exercises

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

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

(iii) $s \in \mathsf{C}_G(\mathrm{ev}(\varphi))$ 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].