13

Logics of Knowledge and Belief

In this chapter we look at how one might represent statements such as "John knows that it is raining," "John believes that it will rain tomorrow," "Mary knows that John believes that it will rain tomorrow" and "It is common knowledge between Mary and John that it is raining."

13.1 The partition model of knowledge

Consider a distributed system, in which multiple processors autonomously performing some joint computation. Of course, the joint nature of the computation means that the processors need to communicate with one another. One set of problems comes about when the communication is error prone. In this case the system analyst may find himself saying something like the following: "Processor A sent the message to processor B. The message may not arrive, and processor A knows this. Furthermore, this is common knowledge, so processor A knows that processor B knows that it (A) knows that if a message was sent it may not arrive." The topic of this chapter is how to make such reasoning precise.

13.1.1 *Muddy children and warring generals*

Often the modeling is done in the context of some stylized problem, with an associated entertaining story. Thus, for example, when we return to the distributed computing application in Section 13.4, rather than speak about computer processors, we will tell the story of two generals who attempt to coordinate among themselves to gang up on a third. For now, however, consider the following less violent story.

> A group of n children enters their house after having played in the mud outside. They are greeted in the hallway by their father, who notices that k of the children have mud on their foreheads. He makes the following announcement, "At least one of you has mud on his forehead." The children can all see each other's foreheads, but not their own. The father then says, "Do any of you know that you have mud on your forehead? If you do, raise

your hand now." No one raises his hand. The father repeats the question, and again no one moves. The father does not give up and keeps repeating the question. After exactly k rounds, all the children with muddy foreheads raise their hands simultaneously.

How can this be? On the face of it only the father's initial statement conveyed new information to the children, and his subsequent questions add nothing. If a child did not have information at the beginning, how could he later on?

Here is an informal argument. Let us start with the simple case in which $k = 1$. In this case the single muddy child knows that all the others are clean, and when the father announces that at least one child is dirty he can conclude that he himself is that child. Note that none of the *other* children know at this point whether or not they are muddy. (After the muddy child raises his hand, however, they do; see next.) Now consider $k = 2$. Imagine that you are one of the two muddy children. After the father's first announcement, you look around the room and see that there is a muddy child other than you. Thus after the father's announcement you do not have enough information to know whether you are muddy (you might be, but it could also be that the other child is the only muddy one). But you note that after the father's first question the other muddy child does not raise his hand. You then realize that you yourself must be muddy as well, or else—based on the reasoning in the $k = 1$ case—that child would have raised his hand. So you raise your hand. Of course, so does the other muddy child.

You could extend this argument to $k = 3, 4, \ldots$, showing in each case that all of the k muddy children raise their hands together after the kth time that the father asks the question. But of course, you would rather have a general theorem that applies to all k. In particular, you might want to prove by induction that after rounds $1, 2, \ldots, k-1$, none of the children know whether they are dirty, but after the next round exactly the muddy children do. However, for this we will need a formal model of "know" that applies in this example.

13.1.2 *Formalizing intuitions about the partition model*

partition model

Definition 13.1.1 (Partition model) *An* (n-agent) partition model *over a language* Σ *is a tuple* $A = (W, \pi, I_1, \ldots, I_n)$*, where:*

- W *is a set of possible worlds;*
- $\pi : \Sigma \mapsto 2^W$ *is an interpretation function that determines which sentences in the languages are true in which worlds; and*
- *each* I_i *denotes a set of possible worlds that are equivalent from the point of view of agent* i*. Formally,* I_i *is a partition of* W*; that is,* $I_i = (W_{i_1}, \ldots, W_{i_r})$ *such that* $W_{i_j} \cap W_{i_k} = \emptyset$ *for all* $j \neq k$*, and* $\cup_{1 \leq j \leq r} W_{i_j} = W$*. We also use the following notation:* $I_i(w) = \{w' \mid w \in W_{i_j} \text{ and } w' \in W_{i_j}\}$*; that is,* $I_i(w)$ *includes all the worlds in the partition of world* w*, according to agent* i*.*

Thus each possible world completely specifies the concrete state of affairs, at least insofar as the language Σ can describe it. For example, in the context of the Muddy Children Puzzle, each possible world will specify precisely which of the children are muddy. The choice of Σ is not critical for current purposes, but for concreteness we will take it to be the languages of propositional logic over some set of primitive propositional symbols.[1] For example, in the context of the muddy children we will assume the primitive propositional symbols `muddy1`, `muddy2`, ..., `muddy`n.

We will use the notation $K_i(\varphi)$ (or simply $K_i\varphi$, when no confusion arises) as "agent i knows that φ."[2] The following defines when a statement is true in a partition model.

Definition 13.1.2 (Logical entailment for partition models) *Let* $A = (W, \pi, I_1, \ldots, I_n)$ *be a partition model over* Σ, *and* $w \in W$. *Then we define the* $\models$ *(logical entailment) relation as follows:*

- *For any* $\varphi \in \Sigma$, *we say that* $A, w \models \varphi$ *if and only if* $w \in \pi(\varphi)$.
- $A, w \models K_i\varphi$ *if and only if for all worlds* w', *if* $w' \in I_i(w)$, *then* $A, w' \models \varphi$.

The first part of the definition gives the intended meaning to the interpretation function π from Definition 13.1.1. The second part states that we can only conclude that agent i knows φ when φ is true in all possible worlds that i considers indistinguishable from the true world.

Let us apply this modeling to the Muddy Children story. Consider the following instance of $n = k = 2$ (i.e., the instance with two children, both muddy). There are four possible worlds, corresponding to each of the children being muddy or not. There are two equivalence relations I_1 and I_2, which allow us to express each of the children's perspectives about which possible worlds can be distinguished. There are two primitive propositional symbols—`muddy1` and `muddy2`. At the outset, before the children see or hear anything, all four worlds form one big equivalence class for each child.

After the children see each other (and can tell apart worlds in which *other* children's state of cleanliness is different) but before the father speaks, the state of knowledge is as illustrated in Figure 13.1. The ovals illustrate the four possible worlds, with the dark oval indicating the true state of the world. The solid boxes indicate the equivalence classes in I_1, and the dashed boxes indicate the equivalence classes in I_2.

Note that in this state of knowledge, in the real world both the sentences K_1`muddy2` and K_2`muddy1` are true (along with, for example, $K_1\neg K_2$`muddy2`). However, neither K_1`muddy1` nor K_2`muddy2` is true in the real world.

1. See Appendix D for a review of propositional logic.
2. The reader familiar with modal logic will note that a partition model is nothing but a special case of a propositional Kripke model with n modalities, in which each of the n accessibility relations is an equivalence relation (i.e., a binary relation that is reflexive, transitive, and symmetric). What is remarkable is that the modalities defined by these accessibility relations correspond well to the notion of knowledge that we have been discussing. This is why we use the modal operator K_i rather than the generic necessity operator $\Box_i$. (The reader unfamiliar with modal logic should ignore this remark; we review modal logic in the next section.)

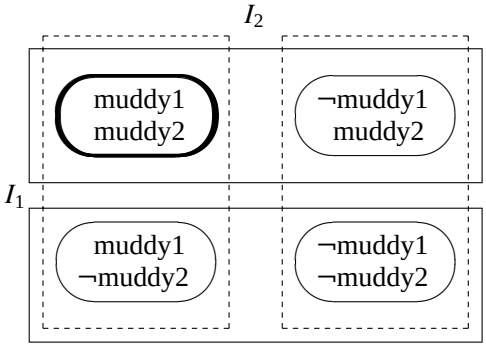

Figure 13.1 Partition model after the children see each other.

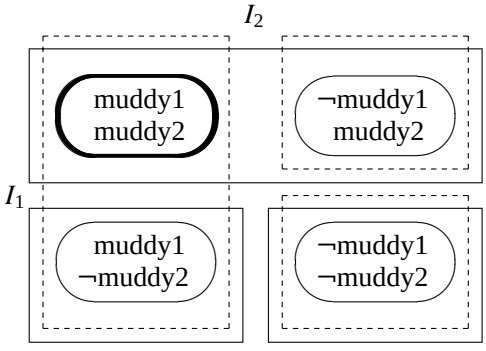

Figure 13.2 Partition model after the father's announcement.

Once the father announces publicly that at least one child is dirty, the world in which neither child is muddy is ruled out. This world then becomes its own partition, leaving the state of knowledge as shown in Figure 13.2.

Thus, were it the case that only one of the children were dirty, at this point he would be able to uniquely identify the real world (and in particular the fact that he was dirty). However, in the real world (where both children are dirty) it is still the case that in this world neither K_1`muddy1` nor K_2`muddy2` holds. However, once the children each observe that the other child does not know his state of cleanliness, the state of knowledge becomes as shown in Figure 13.3. And now indeed both K_1`muddy1` and K_2`muddy2` hold.

Thus, we can reason about knowledge rigorously in terms of partition models. This is a big advance over the previous informal reasoning. But it is also quite cumbersome, especially as these models get larger. Fortunately, we can sometimes reason about such models more concisely using an axiomatic system, which provides a complementary perspective on the same notion of knowledge.

13.2 A detour to modal logic

In order to reason about the partition model of knowledge and later about models of belief and other concepts, we must briefly discuss modal logic. This discussion

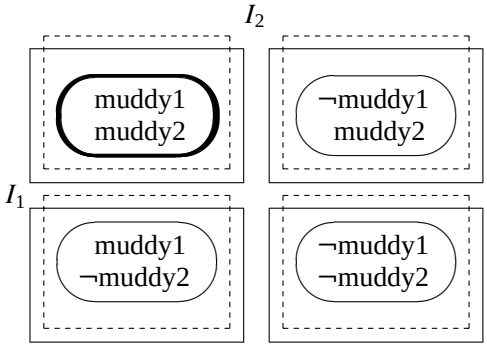

Figure 13.3 Final partition model.

presupposes familiarity with classical logic. For the most part we will consider propositional logic, but we will also make some comments about first-order logic. Both are reviewed in Appendix D.

modal operator

From the syntactic point of view, modal logic augments classical logic with one or more (usually, unary) *modal operators*, and a modal logic is a logic that includes one or more modal operators. The classical notation for a modal operator is $\Box$, often pronounced "necessarily" (and thus $\Box\varphi$ is read as "φ is necessarily true"). The dual modal operator is $\Diamond$, often pronounced "possibly," and is typically related to the necessity operator by the formula $\Diamond\varphi \equiv \neg\Box\neg\varphi$.

What does a modal operator represent, and how is it different from a classical connective such as negation? In general, a modality represents a particular type of judgment regarding a sentence. The default type of judgment, captured in classical logic and not requiring an explicit modal operator, is whether the sentence is true or false. But one might want to capture other sorts of judgments. The original motivation within philosophy for introducing the modal operator is to distinguish between different "strengths of truth." In particular, the wish was to distinguish between accidental truths such as "it is sunny in Palo Alto" (represented, say, by the propositional symbol p), necessary truths such as "either it is sunny in Palo Alto or it is not" ($\Box(p \vee \neg p)$), and possible truths as captured by "it may be sunny in Palo Alto" ($\Diamond p$). A natural hierarchy exists among these three attitudes, with necessary truth implying accidental truth and both implying possible truth. However, the formal machinery has since been used for a variety of other purposes. For example, some logics interpret the modality as quantifying over certain contexts. A case in point are tense logics, or modal temporal logics, in which the context is time. In particular, in some tense logics $\Box\varphi$ is read as "φ is true now and will always be true in the future." We will encounter a similar temporal operator later in the chapter when we discuss robot motion planning. Logics of knowledge and belief read $\Box$ yet differently, as "φ is known" or "φ is believed." These inherently relate the sentence to an *agent*, who is doing the knowing or believing. Indeed, in these logics that interpret the modality in a rather specific way, the $\Box$ notation is usually replaced by other notation that is more indicative of the intended interpretation, but in this section we stick to the generic notation.

Of course, the different interpretations of $\Box$ suggest that there are different modal logics, each endowing $\Box$ with different properties. This indeed is the case. In this section we briefly present the generic framework of modal logic, and in later sections we specialize it to model knowledge, belief, and related notions.

As with any logic, in order to discuss modal logic we need to discuss in turn syntax, semantics, and axiomatics (or proof theory). We concentrate on propositional modal logic, but make some remarks about first-order modal logic at the end.

13.2.1 *Syntax*

The set of sentences in the modal logic with propositional symbols P is the smallest set $\mathcal{L}$ containing P such that if $\varphi, \psi \in \mathcal{L}$ then also $\neg\varphi \in \mathcal{L}$, $\varphi \wedge \psi \in \mathcal{L}$, and $\Box\varphi \in \mathcal{L}$. As usual, other connectives such as $\vee$, $\rightarrow$ and $\equiv$ can be defined in terms of $\wedge$ and $\neg$. In addition, it is common to define the *dual operator* to $\Box$, often denoted $\Diamond$, and pronounced "possibly." It is defined by $\Diamond\varphi \equiv \neg\Box\neg\varphi$, which can be read as "the statement that φ is possibly true is equivalent to the statement that not φ is not necessarily true."

dual operator

13.2.2 *Semantics*

The semantics are defined in terms of *possible-worlds structures*, also called *Kripke structures*. A (single-modality) Kripke structure is a pair (W, R), where W is a collection of (not necessarily distinct) classical propositional models (i.e., models that give a truth value to all sentences that do not contain $\Box$), and R is binary relation on these models. Each $w \in W$ is called a *possible world*. R is called the *accessibility relation*, and sometimes also the *reachability relation* or *alternativeness relation*. It is convenient to think of Kripke structures as directed graphs, with the nodes being the classical models and the arcs representing accessibility.

Kripke structure

possible world

accessibility relation

reachability relation

alternativeness relation

This is where the discussion can start to be related back to the partition model of knowledge. The partition is of course nothing but a binary relation, albeit one with special properties. We will return to these special properties in the next section, but for now let us continue with the generic treatment of modal logic, one that allows for arbitrary accessibility relations.

A truth of a modal sentence is evaluated relative to a particular possible world w in a particular Kripke structure (W, R). (The pair is called a *Kripke model*.) The satisfaction relation is defined recursively as follows:

Kripke model

- $M, w \models p$ if p is true in w, for any primitive proposition p;
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$;
- $M, w \models \neg\varphi$ iff it is not the case that $M, w \models \varphi$;
- $M, w \models \Box\varphi$ iff, for any $w' \in W$ such that $R(w, w')$, it is the case that $M, w' \models \varphi$.

validity

As in classical logic, we overload the $\models$ symbol. In addition to denoting the satisfaction relation, it is used to denote *validity*. $\models \varphi$ means that φ is true in all Kripke models, and, given a class of Kripke models M, $\models_M \varphi$ means that φ is true in all Kripke models within M.

13.2.3 *Axiomatics*

Now that we have a well-defined notion of validity, we can ask whether there exists an axiom system that allows us to derive precisely all the valid sentences, or the valid sentences within a given class. Here we discuss the first question of capturing the sentences that are valid in all Kripke structures; in future sections we discuss specific classes of models of particular interest.

Consider the following axiom system, called the axiom system **K**.

Axiom 13.2.1 (Classical) *All propositional tautologies are valid.*

Axiom 13.2.2 (K) $(\Box\varphi \wedge \Box(\varphi \rightarrow \psi)) \rightarrow \Box\psi$ *is valid.*

Rule 13.2.3 (Modus Ponens) *If both φ and $\varphi \rightarrow \psi$ are valid, infer the validity of ψ.*

Rule 13.2.4 (Necessitation) *From the validity of φ infer the validity of $\Box\varphi$.*

It is not hard to see that this axiom system is *sound*; all the sentences pronounced valid by these axioms and inference rules are indeed true in every Kripke model. What is less obvious, but nonetheless true, is that this is also a *complete* system for the class of all Kripke models; there do not exist additional sentences that are true in all Kripke models. Thus we have the following *representation theorem*:

representation theorem

Theorem 13.2.5 *The system* **K** *is sound and complete for the class of all Kripke models.*

13.2.4 *Modal logics with multiple modal operators*

We have so far discussed a single modal operator and a single accessibility relation corresponding to it. But it is easy to generalize the formulation to include multiple of each. Rather than have a single modal operator $\Box$ we have operators $\Box_1, \Box_2, \ldots, \Box_n$. The set of possible worlds is unchanged, but now we have n accessibility relations: $R_1, R_2, \ldots, R_n$. The last semantic truth condition is changed to:

- For any $i = 1, \ldots, n$, $M, w \models \Box_i\varphi$ iff, for any $w' \in W$ such that $R_i(w, w')$, it is the case that $M, w' \models \varphi$.

Finally, by making similar substitutions in the axiom system **K**, we get a sound and complete axiomatization for modal logic with n modalities. We sometimes denote the system $\mathbf{K_n}$ to emphasize the fact that it contains n modalities. Again, we emphasize that this is soundness and completeness for the class of all n-modality Kripke models. The system remains sound if we restrict the class of

Kripke models under consideration. In general, it is no longer complete, but often we can add axioms and/or inference rules and recapture completeness.

13.2.5 *Remarks about first-order modal logic*

We have so far discussed propositional model logic, that is, modal logic in which the underlying classical logic is propositional. But we can also look at the case in which the underlying logic is a richer first-order language. In such a first-order modal logic we can express sentences such as $\Box\forall x\exists y Father(y, x)$ (read "necessarily, everyone has a father"). We will not discuss first-order in detail, since it will not play a role in what follows. And in fact the technical development is for the most part unremarkable, and it simply mirrors the additional richness of first-order logic as compared to the propositional calculus.

There are however some interesting subtleties, which we point out briefly here.

Barcan formula

The first has to do with the question of whether the so-called *Barcan formula* is valid.

$$\forall x \Box R(x) \rightarrow \Box \forall x R(x)$$

For example, when we interpret $\Box$ as "know," we might ask whether, if for every person it is known individually that the person is honest, it follows that it is known that all people are honest. One can imagine some settings in which the intuitive answer is yes, and others in which it is no. From the technical point of view, the answer depends on the domains of the various models. In first-order modal logic, possible worlds are first-order models. This means in particular that each possible world has a domain, or a set of individuals to which terms in the language are mapped. It is not hard to see that the Barcan formula is valid in the class of first-order Kripke models in which all possible worlds have the same domain, but not in general.

A similar problem has to do with the interaction between modality and equality. If $a = b$, then is it always the case that $\Box(a = b)$? Again, for intuition, consider the knowledge interpretation of $\Box$ and the following famous philosophical example. Suppose there are two objects, the "morning star" and the "evening star." These are two stars that are observed regularly, one in the evening and one in the morning. It so happens these are the same object, namely, the planet Venus. Does it follow that one knows that they are the same object? Intuitively, the answer is no. From the technical point of view, this is because, even if the domains of all possible worlds are the same, the interpretation function which maps language terms to objects might be different in the different worlds. For similar reasons, in general one cannot infer that the world's greatest jazz soprano saxophonist was born in 1897 based on the fact that Sidney Bechet was born that year, since one may not know that Sidney Bechet was indeed the best soprano sax player that ever graced this planet.

13.3 S5: An axiomatic theory of the partition model

We have seen that the axiom system **K** precisely captures the sentences that are valid in the class of all Kripke models. We now return to the partition model of

knowledge, and search for an axiom system that is sound and complete for this more restricted class of Kripke models. Since it is a smaller class, it reasonable to expect that it will be captured by an expanded set of axioms. This is indeed the case.

We start with system **K**, but now use the K_i to denote the modal operators rather than $\Box_i$, representing the interpretation we have in mind.[3]

Axiom 13.3.1 (Classical) *All propositional tautologies are valid.*

Axiom 13.3.2 (K) $(K_i\varphi \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i\psi$

Rule 13.3.3 (Modus Ponens) *From* φ *and* $\varphi \rightarrow \psi$ *infer* ψ.

Rule 13.3.4 (Necessitation) *From* φ *infer* $K_i\varphi$.

Note that the generic axiom **K** takes on a special meaning under this interpretation of the modal operator—it states that an agent knows all of the tautological consequences of that knowledge. Thus, if you know that it is after midnight, and if you know that the president is always asleep after midnight, then you know that the president is asleep. Less plausibly, if you know the Peano axioms, you know all the theorems of arithmetic. For this reason, this property is sometimes called *logical omniscience*. One could argue that this property is too strong, that it is too much to expect of an agent. After all, humans do not seem to know the full consequences of all of their knowledge. Similar criticisms can be levied also against the other axioms we discuss below. This is an important topic, and we returned to it later. However, it is a fact that the idealized notion of "knowledge" defined by the partition model has the property of logical omniscience. So to the extent that the partition model is useful (and it is, as we discuss later in the chapter), one must learn to live with some of the strong properties it induces on "knowledge."

logical omniscience

Other properties of knowledge do not hold for general Kripke structures, and are thus not derivable in the **K** axiom system. One of these properties is *consistency*, which states that an agent cannot know a contradiction. The following axiom, called axiom **D** for historical reasons, captures this property:

consistency

Axiom 13.3.5 (D) $\neg K_i(p \wedge \neg p)$.

Note that this axiom cannot be inferred from the axiom system **K** and is thus not valid in the class of all n-Kripke structures. It is, however, valid in the more restricted class of *serial models*, in which each accessibility is serial. (A binary relation X over domain Y is serial if and only if $\forall y \in Y\ \exists y' \in Y$ such that $(y, y') \in X$.) Indeed, as we shall see, it is not only sound but also complete. The axiom system obtained by adding axiom **D** to the axiom system **K** is called axiom system **KD**.

serial accessibility relation

Another property that holds for our current notion of knowledge is that of *veridity*; it is impossible for an agent to know something that is not actually true.

veridity

3. We stick to this notation for historical accuracy, but the reader should not be confused by it. The axiomatic system **K** and the axiom **K** that gives rise to the name have nothing to do with the syntactic symbol K_i we use to describe the knowledge of an agent.

Indeed, this is often taken to be the property that distinguishes knowledge from other informational attitudes, such as belief. We can express this property with the so-called axiom **T**:

Axiom 13.3.6 (T) $K_i\varphi \rightarrow \varphi$.

This axiom also cannot be inferred from the axiom system **KD**. Again, the class of Kripke structures for which axiom **T** is sound and complete can be defined succinctly—it consists of the Kripke structures in which each accessibility relation is *reflexive*. (A binary relation X over domain Y is reflexive if and only if $\forall y \in Y, (y, y) \in X$. Note that $y = y'$ is allowed.) By adding axiom **T** to the system **KD** we get the axiom system **KDT**. However, in this system the axioms are no longer independent; the axiom system **KDT** is equivalent to the axiom system **KT**, since the axiom **D** can be derived from the remaining axioms and the inference rules. Indeed, this follows easily from the completeness properties of the individual axioms; every reflexive relation is trivially also serial.

reflexive accessibility relation

There are two additional properties of knowledge induced by the partition model which are not captured by the axioms discussed thus far. They both have to do with the introspective capabilities of a given agent, or the nesting of the knowledge operation. We first consider *positive introspection*, the property that, when an agent knows something, he knows that he knows it. It is expressed by the following axiom, historically called axiom **4**:

positive introspection

Axiom 13.3.7 (4) $K_i\varphi \rightarrow K_iK_i\varphi$.

Again, it does not follow from the other axioms discussed so far. The class of Kripke structures for which axiom **4** is sound and complete consists of the structures in which each accessibility relation is transitive. (A binary relation X is transitive if and only if for all $y, y', y'' \in Y$ it is the case that if $(y, y') \in X$ and $(y', y'') \in X$ then $(y, y'') \in X$.)

The last property of knowledge we will consider is *negative introspection*. This is quite similar to positive introspection, but here we are concerned that if an agent does not know something, then he knows that he does not know it. We express this with the following axiom, called axiom **5**:

negative introspection

Axiom 13.3.8 (5) $\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$.

Again, it does not follow from the other axioms. Consider now the class of Kripke structures in which axiom **5** is sound and complete. Informally speaking, we want to ensure that if two worlds are accessible from the current world, then they are also accessible from each other. Formally, we say that the accessibility relation must be *Euclidean*. (A binary relation X over domain Y is Euclidean if and only if for all $y, y', y'' \in Y$ it is the case that if $(y, y') \in X$ and $(y, y'') \in X$ then $(y', y'') \in X$.)

Euclidean accessibility relation

At this point the reader may feel confused. After all, we started with a very simple class of Kripke structures—partition models—in which each accessibility relation is a simple equivalence relation. Then, in our pursuit of axioms that capture the notion of knowledge defined in partition models, we have looked at increasingly baroque properties of accessibility relations and associated axioms, as summarized in Table 13.1.

Name	Axiom	Accessibility Relation
Axiom **K**	$(K_i(\varphi) \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i(\psi)$	NA
Axiom **D**	$\neg K_i(p \wedge \neg p)$	Serial
Axiom **T**	$K_i\varphi \rightarrow \varphi$	Reflexive
Axiom **4**	$K_i\varphi \rightarrow K_iK_i\varphi$	Transitive
Axiom **5**	$\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$	Euclidean

Table 13.1 Axioms and corresponding constraints on the accessibility relation.

What do these complicated properties have to do with the simple partition models that started this discussion? The answer lies in the following observation.

Proposition 13.3.9 *A binary relation is an equivalence relation if and only if it is reflexive, transitive, and Euclidean.*

Indeed, the system **KT45** (which results from adding to the axiom system **K** all the axioms **T**, **4**, and **5**), exactly captures the properties of knowledge as defined by the partition model. System **KT45** is also known by another, more common name—the **S5** axiom system. **S5** is both sound and complete for the class of all partition models. However, we are able to state an even more general theorem, which will serve us well later when we discuss moving from knowledge to belief.

Theorem 13.3.10 *Let* $\mathbf{X}$ *be a subset of* $\{\mathbf{D}, \mathbf{T}, \mathbf{4}, \mathbf{5}\}$ *and let* $\mathcal{X}$ *be the corresponding subset of* $\{$*serial, reflexive, transitive, Euclidean*$\}$. *Then* $\mathbf{K} \cup \mathbf{X}$ *(which is the basic axiom system* **K** *with the appropriate subset of axioms added) is a sound and complete axiomatization of* K_i *for the class of Kripke structures whose accessibility relations satisfy* $\mathcal{X}$.

13.4 Common knowledge, and an application to distributed systems

Earlier we discussed the domain of distributed systems. The following example illustrates the sort of reasoning one would like to perform in that context. (In case this problem does not sound like distributed computing to you, imagine that the generals are computer processes, which are trying to communicate reliably over a faulty communication line.)

> Two generals standing on opposing hilltops are trying to communicate in order to coordinate an attack on a third general, whose army sits in the valley between them. The two generals are communicating via messengers who must travel across enemy lines to deliver their messages. Any messenger carries the risk of being caught, in which case the message is lost. (Alas, the fate of the messenger is of no concern in this story.) Each of the two generals wants to attack, but only if the other does; if they both attack they will win, but either one will lose if he attacks alone. Given this context,

what protocol can the generals establish that will ensure that they attack simultaneously?

You might imagine the following naive communication protocol. The protocol for the first general, S, is to send an "attack tomorrow" message to general R and to keep sending this message repeatedly until he receives an acknowledgment that the message was received. The protocol for the second general, R, is to do nothing until he receives the message from S, and then send a single acknowledgment message back to general S. The question is whether the agents can have a plan of attack based on this protocol, which always guarantees—or, less ambitiously, can sometimes guarantee—that they attack simultaneously. And if not, can a different protocol achieve this guarantee?

Clearly, it would be useful to reason about this scenario rigorously. It seems intuitive that the formal notion of "knowledge" should apply here, but the question is precisely how. In particular, what is the knowledge condition that must be attained in order to ensure a coordinated attack?

To apply the partition model of knowledge we need to first define the possible worlds. We will do this by first defining the *local state* of each agent (i.e., each general); together the two local states will form a *global state*. To reason about the naive protocol, we will have the local state for S represent two binary pieces of information: whether or not an "attack" message was sent and whether or not an acknowledgment was received. The local state for R will also represent two binary pieces of information: whether or not an "attack" message was received and whether or not an acknowledgment message was sent. Thus we end up with the four possible local states for each general—(0, 0), (0, 1), (1, 0), and (1, 1)—and thus sixteen possible global states.

We are now ready to define the possible worlds of our model. The initial global state is well defined—by convention, we call it $\langle(0, 0), (0, 0)\rangle$. It will then evolve based on two factors—the dictates of the protocol, and the nondeterministic effect of nature (which decides whether a message is received or not). Thus, among all the possible sequences of global states, only some are legal according to the protocol. We will call any finite prefix of such a legal sequence of global states a *history*. These histories will be the possible worlds in our model.

history

For example, consider the following possible sequence of events, given the naive protocol: S sends an "attack" message to R, R receives the message and sends an acknowledgment to S, and S receives the acknowledgment. The history corresponding to this scenario is

$$\langle(0, 0), (0, 0)\rangle, \langle(1, 0), (1, 0)\rangle, \langle(1, 1), (1, 1)\rangle.$$

The structure of our possible worlds suggests a natural definition of the partition associated with each of the agents. We will say that two histories are in the same equivalence class of agent i ($i \in \{S, R\}$) if their respective final global states have identical local state for agent i. That is, history

$$\langle(0, 0), (0, 0)\rangle, \langle x_{S,1}, x_{R,1}\rangle, \ldots, \langle x_{S,k}, x_{R,k}\rangle$$

is indistinguishable in the eyes of agent i from history

$$\langle (0,0),(0,0)\rangle, \langle y_{S,1}, y_{R,1}\rangle, \ldots, \langle y_{S,l}, y_{R,l}\rangle$$

if and only if $x_{i,k} = y_{i,l}$. Note a very satisfying aspect of this definition; the accessibility relation, which in general is thought of as an abstract notion, is given here a concrete interpretation in terms of local and global states.

We can now reason formally about the naive protocol. Consider again the possible history mentioned above:

$$\langle (0,0),(0,0)\rangle, \langle (1,0),(1,0)\rangle, \langle (1,1),(1,1)\rangle.$$

The reader can verify that in this possible world the following sentences are true: $K_S\texttt{attack}$, $K_R\texttt{attack}$, $K_SK_R\texttt{attack}$. However, it is also the case that this world satisfies $\neg K_RK_SK_R\texttt{attack}$. Intuitively speaking, R does not know that S knows that R knows that S intends to attack, since for all R knows its acknowledgment could have been lost. Thus R cannot proceed to attack; he reasons that if indeed the acknowledgement was lost then S will not dare attack.

This of course suggests a fix—have S acknowledge R's acknowledgement. To accommodate this we would need to augment each of the local states with another binary variable, representing for S whether the second acknowledgement was sent and for R whether it was received. However it is not hard to see that this also has a flaw; assuming S's second acknowledgement indeed goes through, in the resulting history we will have $K_RK_SK_R\texttt{attack}$ hold, but not $K_SK_RK_SK_R\texttt{attack}$. Can this be fixed, and what is the general knowledge condition we must aim for?

common knowledge

It turns out that the required condition is what is known as *common knowledge*. This is a very intuitive notion, and we define it in two steps. We first define what it means that *everybody knows* a particular sentence to be true. To represent this we use the symbol E_G, where G is a particular group of agents. The "everybody knows" operator has the same syntactic rules as the knowledge operator. As one might expect, the semantics can be defined easily in terms of the basic knowledge operator. We define the semantics as follows.

Definition 13.4.1 ("Everyone knows") *Let M be a Kripke structure, w be a possible world in M, G be a group of agents, and φ be a sentence of modal logic. Then $M, w \models E_G\varphi$ if and only if $M, w' \models \varphi$ for all $w' \in \cup_{i\in G} I_i(w)$. (Equivalently, we can require that $\forall i \in G$ it is the case that $M, w \models K_i\varphi$.)*

In other words, everybody knows a sentence when the sentence is true in all of the worlds that are considered possible in the current world by any agent in the group.

Using this concept we can define the notion of *common knowledge*, or, as it is sometimes called lightheartedly, "what any fool knows." If "any fool" knows something, then we can assume that everybody knows it, and everybody knows that everybody knows it, and so on. An example from the real world might be the assumption we use when driving a car that all of the other drivers on the road also know the rules of the road, and that they know that we know the rules of the road, and that they know that we know that they know them, and so on. We

require an infinite series of "everybody knows" in order to capture this intuition. For this reason we use the following recursive definition.

Definition 13.4.2 (Common knowledge) *Let M be a Kripke structure, w be a possible world in M, G be a group of agents, and φ be a sentence of modal logic. Then $M, w \models C_G\varphi$ if and only if $M, w \models E_G(\varphi \wedge C_G\varphi)$.*

fixed-point axiom

In other words, a sentence is common knowledge if everybody knows it and knows that it is common knowledge. This formula is called the *fixed-point axiom*, since $C_G\varphi$ can be viewed as the fixed-point solution of the equation $f(x) = E_G(\varphi \wedge f(x))$. Fixed-point definitions are notoriously hard to understand intuitively. Fortunately, we can give alternative characterizations of C_G. First,we can give a direct semantic definition.

Theorem 13.4.3 *Let M be a Kripke structure, w be a possible world in M, G be a group of agents, and φ be a sentence of modal logic. Then $M, w \models C_G\varphi$ if and only if $M, w' \models \varphi$ for every sequence of possible worlds $(w = w_0, w_1, \ldots, w_n) = w'$ for which the following holds: for every $0 \leq i < n$ there exists an agent $j \in G$ such that $w_{i+1} \in I_j(w_i)$.*

Second, it is worth noting that our **S5** axiomatic system can also be enriched to provide a sound and complete axiomatization of C_G. It turns out that what are needed are two additional axioms and one new inference rule.

Axiom 13.4.4 (A3) $E_G\varphi \leftrightarrow \bigwedge_{i \in G} K_i\varphi$

Axiom 13.4.5 (A4) $C_G\varphi \rightarrow E_G(\varphi \wedge C_G\varphi)$

Rule 13.4.6 (R3) *From $\varphi \rightarrow E_G(\psi \wedge \varphi)$ infer $\varphi \rightarrow C_G\psi$*

This last inference rule is a form of an *induction rule*.

Armed with this notion of common knowledge, we return to our warring generals. It turns out that, in a precise sense, whenever any communication protocol guarantees a coordinated attack in a particular history, in that history it must achieve common knowledge between the two generals that an attack is about to happen. It is not hard to see that no finite exchange of acknowledgments will ever lead to such common knowledge. And thus it follows that there is *no* communication protocol that solves the Coordinated Attack problem, at least not as the problem is stated here.

13.5 Doing time, and an application to robotics

We now move to another application of reasoning about knowledge, in robotics. The domain of robotics is characterized by *uncertainty*; a robot receives uncertain readings from its input devices, and its motor controls produce imprecise motion. Such uncertainty is not necessarily the end of the world, so long as its magnitude is not too large for the task at hand, and that the robot can reason about it effectively. We will explicitly consider the task of robot motion planning under uncertainty, and in particular the question of how a robot knows to stop despite

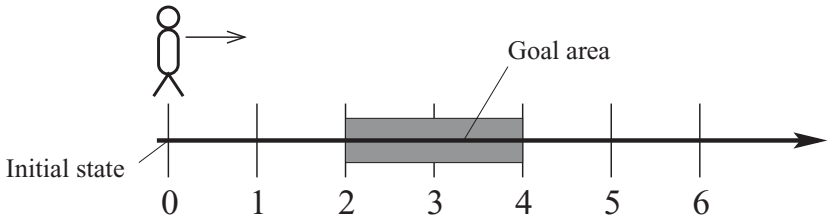

Figure 13.4 A one-dimensional robot motion problem.

having an imprecise sensor and perhaps also motion controller. We first discuss the single-robot case, where we see the power of the knowledge abstraction. We then move to the multiagent case, where we show the importance of each agent being able to model the other agents in the system.

13.5.1 *Termination conditions for motion planning*

Imagine a point robot moving in one dimension along the positive reals from the origin to the right, with the intention of reaching the interval [2, 4], the goal region (see Figure 13.4). The robot is moving at a fixed finite velocity, so there is no question about whether it will reach the destination; the question is whether the robot will know when to stop. Assume the robot has a position sensor that is inaccurate; if the true position is L, the sensor will return any value in the interval $[L - 1, L + 1]$. We assume that time is continuous and that the sensor provides readings continuously, but that the reading values are not necessarily continuous. We are looking for a termination condition—a predicate on the readings such that as soon as it evaluates to "true" the robot will stop. What is a reasonable termination condition?

Consider the termination condition "$R = 3$" where R is the current reading. This is a "safe" condition, in the sense that if that is the reading then the robot is guaranteed to be in the goal region. The problem is that, because of the discontinuity of reading values, the robot may never get that reading. In other words, this termination condition is sound but not complete. Conversely, the condition "$2 \leq R \leq 4$" is complete but not sound. Is there a sound and complete termination condition?

On reflection it is not hard to see that "$R \geq 3$" is such a condition. Although there are locations outside the goal region that can give rise to readings that satisfy the predicate (e.g., in location 10 the reading necessarily does), what matters is the first time the robot encounters that reading. Given its starting location and its motion trajectory, the first time can be no earlier than 2 and no later than 4.

Let us now turn to a slightly more complex, two-dimensional robotic motion planning problem, depicted in Figure 13.5. A point robot is given a command to move in a certain direction, and its goal is to arrive at the depicted rectangular goal region. As in the previous example, its sensor is error prone; it returns a reading that is within ρ from the true location (i.e., the uncertainty is captured by a disk of radius ρ). Unlike the previous example, the robot's motion is also subject to

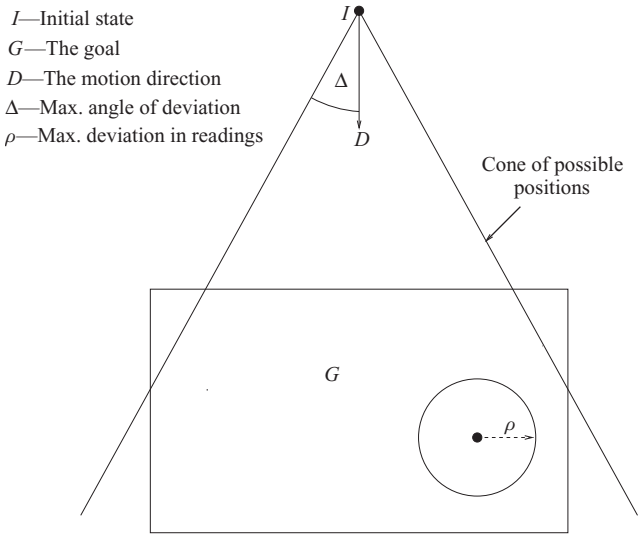

Figure 13.5 A two-dimensional robot motion problem.

error; given a command to move in a certain direction, the robot's actual motion can deviate up to Δ degrees from the prescribed direction. It is again assumed that when the robot is given the command "Go in direction D" it moves at some finite velocity, always moving within Δ degrees from that direction, but that the deviation is not consistent and can change arbitrarily within the tolerance Δ. It is not hard to see that the space of locations reachable by the robot is described by a cone whose angle is 2Δ, as depicted in Figure 13.5. Again we ask, what is a good termination condition for the robot? One candidate is a rectangle inside the goal region whose sides are Δ away from the corresponding side of the goal rectangle. Clearly this is a sound condition, but not a complete one. Figure 13.6 depicts another candidate, termed the "naive termination condition"—all the sensor readings in the region with the dashed boundary. Although it is now less obvious, this termination condition is also sound—but, again, not complete. In contrast, Figure 13.7 shows a termination condition that is both sound and complete.

Consider now the two sound and complete termination conditions identified for the two problems, the one-dimensional one and the two-dimensional. Geometrically, they seem to bear no similarity, and yet one's intuition might be that they embody the same principle. Can one articulate this principle, and thus perhaps later on apply it to yet other problems? To do so we abstract to what is sometimes called the knowledge level. Rather than speak about the particular geometry of the situation, let us reason more abstractly about the knowledge available to the robot. Not surprisingly, we choose to do it using possible-worlds semantics and specifically the S5 model of knowledge. Define a history of the robot as a mapping from time to both location and sensor reading. That is, a history tells us where the robot has been at any point in time and its sensor reading at that time. Clearly, every motion planning problem—including both the ones presented here—defines a set of legal histories. For every motion planning

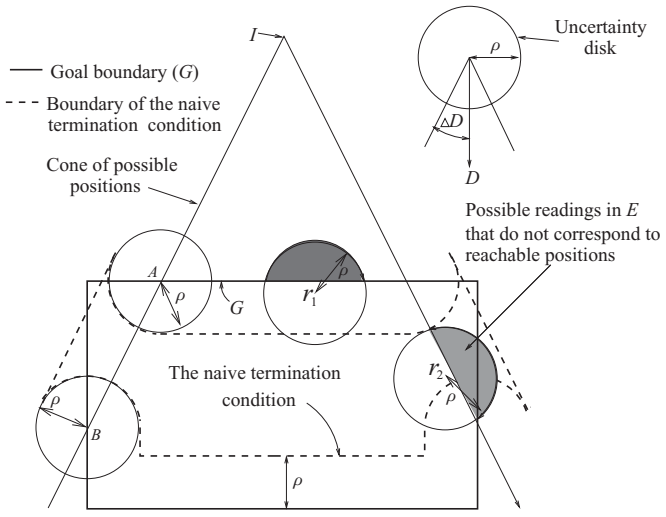

Figure 13.6 A sound but incomplete termination condition for the two-dimensional robot motion problem.

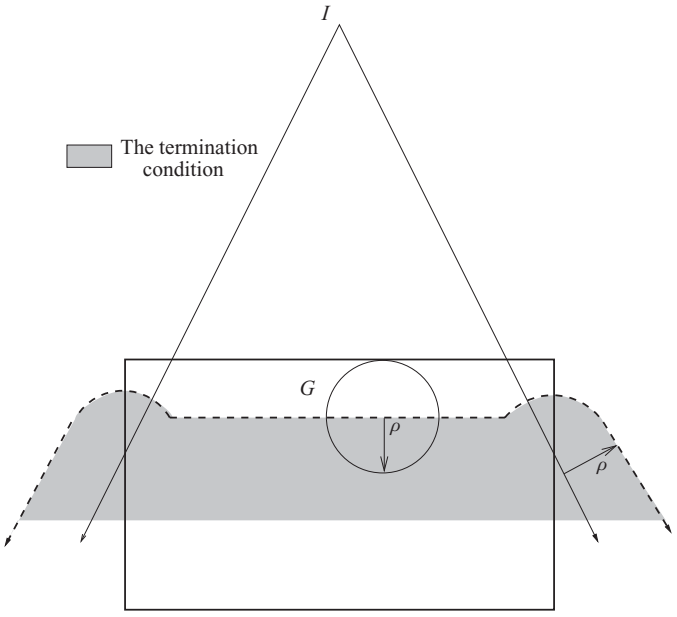

Figure 13.7 A sound and complete termination condition for the two-dimensional robot motion problem.

problem, our set of possible worlds will consist of pairs (h, t), where h is a legal history and t is a time point. We will say that (h, t) is accessible from (h', t') if the sensor reading in h at t is identical to the sensor reading in h' at t'. Clearly, this defines an equivalence relation. We now need a language to speak about such possible-worlds structures. As usual, we will use the K operator to denote knowledge, that is, truth in all accessible worlds. We will also use a temporal operator $\Diamond$ whose meaning will be "sometime in the past."

Specifically, $h, t \models \Diamond\varphi$ holds at h, t just in case there exists a $t' \leq t$ such that $h, t' \models \varphi$.[4] Armed with this, we are ready to provide a knowledge–level condition that captures both the specific geometric ones given earlier. It is deceptively simple. Let g be the formula meaning "I am in the goal region" (which is of course instantiated differently in the two examples). Then the sound and complete termination condition is simply $K\Diamond g$, reading informally "I know that either I am in the goal region now, or else sometime in the past I was." It turns out that this simple termination condition is not only sound and complete for these two examples, but is so also in a much larger set of examples, and is furthermore optimal: it causes the robot to stop no later than any other sound and complete termination condition.

13.5.2 *Coordinating robots*

We now extend this discussion to a multiagent setting. Add to the first example a second robot that must coordinate with the first robot (number the robots 1 and 2). Specifically, robot 2 must take an action at the exact time (and never before) robot 1 stops in the goal area. For concreteness, think of the robots as teammates in the annual robotic competition, where robotic soccer teams compete against each other. Robot 2 needs to pass the ball to robot 1 in front of the opposing goal. Robot 1 must be stopped to receive the pass, but the pass must be completed as soon it stops, or else the defense will have time to react. The robots must coordinate on the pass without the help of communication. Instead, we provide robot 2 with a sensor of robot 1's position. Let R_1 be the sensor reading of the first robot, R_2 be that of the second, and p the true position of robot 1.

The sound and complete termination condition that we seek in this setting is actually a conjunction of termination conditions for the two robots. Soundness means that robot 1 never stops outside the goal area and robot 2 never takes its action when robot 1 is not stopped in the goal area. Completeness obviously means that the robots eventually coordinate on their respective actions. As in the example of the warring generals, the two robots need common knowledge in order to coordinate. For example, if both robots know that robot 1 is in the goal but robot 1 does not know that robot 2 knows this, then robot 1 cannot stop, because robot 2 may not know to pass the ball at that point in time. Thus, the sound and complete termination condition of $K(\Diamond g)$ in the single robot setting becomes $C_{1,2}(\Diamond g)$ in the current setting, where g now means "Robot 1 is in the goal." That is, when this fact becomes common knowledge for the first time, the robots take their respective actions.

So far, we have left the sensor of robot 2 unspecified. In the first instance we analyze, let robot 2 be equipped with a sensor identical to robot 1's. That is, the sensor returns the position of robot 1 within the interval $[L - 1, L + 1]$, and the noise of the sensor is deterministic so that this value is exactly equal to the value observed by robot 1. Further, assume that this setting is common knowledge between the two robots. We can formalize the setting as a partition

4. This operator is taken from the realm of temporal logic, where it usually appears in conjunction of other modalities; however, those are not required for our example.

model in which a possible world (which was (h, t) above) is a tuple $\langle p, R_1, R_2 \rangle$. The partitions of the two robots define equivalence classes based on the sensor of that agent. That is, $P_i(\langle p, R_1, R_2 \rangle) = \{\langle p', R_1', R_2' \rangle | (R_i = R_i')\}$, for $i = 1, 2$. The common knowledge of the properties of the sensors reduces the space of possible worlds to all $\langle p, R_1, R_2 \rangle$ such that $|p - R_1| \leq 1$ and $R_1 = R_2$. The latter property implies that the partition structures for the two agents are identical.

We can now analyze this partition model to find our answer. From the single robot case, we know that $R_1 \geq 3 \rightarrow K_1(\Diamond g)$. This statement can be verified easily using the partition structure. Recall the semantic definition of common knowledge: the proposition must be true in the last world of every sequence of possible worlds that starts at the current world and only transitions to a world that is in the partition of an agent. Because the partition structures are identical, starting from a world in which $R_1 \geq 3$, we cannot transition out of the identical partition we are in for both agents, and thus $C_{1,2}(\Diamond g)$ holds. Therefore, the robots can coordinate by using the same rule as the previous section: robot 1 stops when $R_1 \geq 3$ is first true, and robot 2 takes its action when $R_2 \geq 3$ is first true. Since any lower sensor reading does not permit either robot to know $\Diamond g$, it is also the case that this rule is optimal.

One year later, we are preparing for the next robotic competition. We have received enough funding to purchase a perfect sensor (one that always returns the true position of robot 1). We replace the sensor for robot 2 with this sensor, but we do not have enough money left over to buy either a new sensor for robot 1 or a means of communication between the two robots. Still, we are optimistic that our improved equipment will allow us to fare better this year by improving the termination condition so that the pass can be completed earlier. Since we have different common knowledge about the sensors, we have a different set of possible worlds. In this instance, each possible world $\langle p, R_1, R_2 \rangle$ must satisfy $p = R_2$ instead of $R_1 = R_2$ while also satisfying $|p - R_1| \leq 1$ as before. This change causes the robots to no longer have identical partition structures. Analyzing the new structure, we quickly realize that not only can we not improve the termination condition, but also that our old rule no longer works. As soon as $R_2 \geq 3$ becomes true, we have both $K_1(\Diamond g)$ and $K_2(\Diamond g)$. However, in the partition for robot 2 in which $R_2 = 3$, we find possible worlds in which $R_1 < 3$. In these worlds, robot 1 does not know that it is in the goal. Thus, we have $\neg K_2(K_1(\Diamond g))$, which means that robot 2 cannot take the action because robot 1 might not stop to receive the pass.

Suppose we try to salvage the situation by implementing a later termination condition. This idea obviously fails, because if robot 2 instead waits for R_2 to be a value greater than 3, then robot 1 may have already stopped. However, our problems run deeper. Even if we only require that the robots coordinate at some time after robot 1 enters the goal area (implicitly extending the goal to the range $[2, \infty]$), we can never find a sound and complete termination condition. Common knowledge of $\Diamond g$ is impossible to achieve, because, from any state of the world $\langle p, R_1, R_2 \rangle$, we can follow a path of possible worlds (alternating between the robot whose partition we transition in) until we get to a world in which robot 1 is not in the goal. For example, consider the world $\langle 5, 6, 5 \rangle$. Robot 2 considers it possible that robot 1 observes a value of 4 (in world $\langle 5, 4, 5 \rangle$). In that world, robot 1 would consider it possible that the true world is 3 (in world $\langle 3, 4, 3 \rangle$).

We can then make two more transitions to $\langle 3, 2, 3\rangle$ and then to $\langle 1, 2, 1\rangle$, in which robot 1 is not in the goal. Thus, we have $\neg K_2(K_1(K_2(K_1(\Diamond g))))$. Since we can obviously make a similar argument for any world with a finite value for p, there does not exist a world that implies $C_{1,2}(\Diamond g)$.

This example illustrates how, when coordination is necessary, knowledge of the knowledge of the other agents can be more important than objective knowledge of state of the world. In fact, in many cases it takes common knowledge to achieve common knowledge. When robot 2 had the flawed sensor, we needed common knowledge of the fact that the sensors are identically flawed (which was encoded in the space of possible worlds) in order for the robots to achieve common knowledge of $\Diamond g$. Also, the fact that the agents have to coordinate is a key restriction in this setting. If instead, we only needed robot 1 to stop in the goal and robot 2 to take its action at some point after robot 1 stopped, then all we need is $K_1(\Diamond g)$ and $K_2(K_1(\Diamond g))$. This is achieved by the termination conditions of $R_1 \geq 3$ for robot 1 and $R_2 \geq 4$ for robot 2.

13.6 From knowledge to belief

We have so far discussed a particular informational attitude, namely "knowledge," but there are others. In this section we discuss "belief," and in the next section we will mention a third—"certainty."

Like knowledge, belief is a mental attitude, and concerns an agent's view of different state of affairs. Indeed, we will model belief using Kripke structures (i.e., possible worlds with binary relations defined on them). However, intuition tells us that belief has different properties from those of knowledge, which suggests that these Kripke structures should be different from the partition models capturing knowledge.

We will return to the semantic model of belief shortly, but it is perhaps easiest to make the transition from knowledge to belief via the axiomatic system. Recall that in the case of knowledge we had the veridity axiom, **T**: $K_i\varphi \rightarrow \varphi$. Suppose we take the **S5** system and simply drop this axiom—what do we get? Hidden in this simple question is a subtlety, it turns out. Recall that **S5** was shorthand for the system **KDT45**. Recall also that **KDT45** was logically equivalent to **KT45**. However, **KD45** is *not* logically equivalent to **K45**; axiom **D** is not derivable without axiom **T** (why?). It turns out that both **KD45** and **K45** have been put forward as a logic of idealized belief, and in fact both have been called *weak* **S5**. There are good reasons, however, to stick to **KD45**, which we will do henceforth.

The standard logic of belief **KD45** therefore consists of the following axioms; note that we change the notation K_i to B_i to reflect the fact that we are modeling belief rather than knowledge.

Axiom 13.6.1 (K) $(B_i\varphi \wedge B_i(\varphi \rightarrow \psi)) \rightarrow B_i\psi$

Axiom 13.6.2 (D) $\neg B_i(p \wedge \neg p)$

Axiom 13.6.3 (4) $B_i\varphi \rightarrow B_i B_i\varphi$

Axiom 13.6.4 (5) $\neg B_i\varphi \rightarrow B_i\neg B_i\varphi$

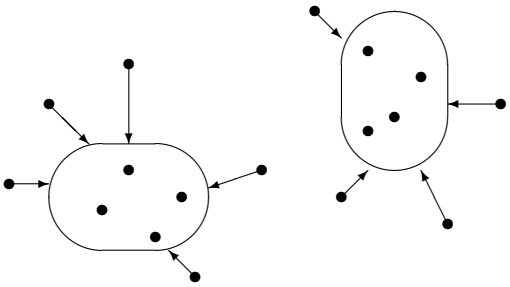

Figure 13.8 Graphical depiction of a quasi-partition structure.

These axioms—logical omniscience, consistency, positive and negative introspection—play the same roles as in the logic of knowledge, as do the two inference rules, *Modus Ponens* and *Necessitation*.

Modus Ponens

Necessitation

The next natural question is whether we can complement this axiomatic theory of belief with a semantic one. That is, is there a class of Kripke structures for which this axiomatization is sound and complete, just as **S5** is sound and complete for the class of partition models? The theorem in Section 13.3 has already provided us the answer; it is the class of Kripke structures in which the accessibility relations are serial, transitive, and Euclidean. Just as in the case of knowledge, there is a relatively simple way to understand this class, although it is not as simple as the class of partitions. As shown in Figure 13.8, we can envision such a structure as composed of several clusters of worlds, each of which is completely connected internally. In addition there are possibly some singleton worlds, each of which is connected only to all the worlds within exactly one cluster. We call such a model a *quasi-partition*.

quasi-partition

We saw that it was useful to define the notion of common knowledge. Can we similarly define *common belief*, and would it be a useful one? The answers are yes, and maybe. We can certainly define common belief by mirroring the definitions for knowledge. However, the resulting notion will necessarily be weaker. In particular, it is not hard to verify the validity of the following sentence:

common belief

$$K_i C_G \varphi \equiv C_G \varphi.$$

This equivalence breaks down in the case of belief and common belief.

13.7 Combining knowledge and belief (and revisiting knowledge)

Up until this point we have discussed how to model knowledge and belief separately from each other. But of course we may well want to combine the two, so that we can formalize sentences such as "if Bob knows that Alice believes it is raining, then Alice knows that Bob knows it." Indeed, it is easy enough to just merge the languages of knowledge and belief, so as to allow the sentence:

$$K_B B_A \texttt{rain} \rightarrow K_A K_B B_A \texttt{rain}.$$

Furthermore, there is no difficulty in merging the semantic structures and having two sets of accessibility relations over the possible worlds—partition models representing the knowledge of the agents and quasi-partition models representing their beliefs. Finally, we can merge the axiom systems of knowledge and belief and obtain a sound and complete axiomatization of merged structures.

However, doing just these merges, while preserving the individual properties of knowledge and belief, will not capture any interaction between them. And we do have some strong intuitions about such interactions. For example, according to the intuition of most people, knowledge implies belief. That is, the following sentence ought to be valid:

$$K_i\varphi \rightarrow B_i\varphi.$$

This sentence is not valid in the class of all merged Kripke structures defined earlier. This means that we must introduce further restrictions on these models in order to capture this property, and any other property about the interaction between knowledge and belief that we care about.

We will introduce a particular way of tying the two notions together, which has several conceptual and technical advantages. It will force us, however, to reconsider some of the assumptions we have made about knowledge and belief thus far, and to enter into philosophical discussion more deeply than we have heretofore.

defeasible

To begin, we should distinguish two types of belief. Both of them are distinguished from knowledge in that they are *defeasible*; the agent may believe something false. However, in one version of belief the believing agent is aware of this defeasibility, while in the other it is not. We will call the first type of belief "mere belief" (or sometimes "belief" for short) and the second type of belief "certainty." An agent who is certain of a fact will not admit that he might be wrong; to him, his beliefs look like knowledge. It is only *another* agent who might label his beliefs as only that and deny them the status of knowledge. Imagine that John is certain that the bank is open, but in fact the bank is closed. If you were to ask John, he would tell you that the bank is open. If you pressed him with "Are you sure?" he would answer "What do you mean 'am I sure,' I *know* that it is open." But of course this cannot be knowledge, since it is false. In contrast, you can imagine that John is pretty sure that the bank is open, sufficiently so to make substantial plans based on this belief. In this case if you were to ask John "Is the bank open?" he might answer "I believe so, but I am not sure." In this case John would be the first to admit that this is a mere belief on his part, not knowledge.[5]

5. A reader seeped in the Bayesian methodology would at this point be strongly tempted to protest that if John is not sure he should quantify his belief probabilistically, rather than make a qualitative statement. We will indeed discuss probabilistic beliefs in Section 14.1. But one should not discard qualitative beliefs casually. Not only can psychological and commonsense arguments be made on behalf of this notion, but in fact the notion plays an important role in game theory, which is for the most part Bayesian in nature. We return to this later in this chapter, when we discuss *belief revision* in Section 14.2.1.

From a technical point of view, we can split the B_i operator into two versions, B_i^m for "mere belief" and B_i^c for certainty. Even before entering into a formal account of such operators, we expect that the sentence $B_i^c\varphi \rightarrow B_i^c K_i\varphi$ will be valid, but the sentence $B_i^m\varphi \rightarrow B_i^m K_i\varphi$ will not.

This distinction between certainty and mere belief also calls into question some of our assumptions about knowledge, in particular the negative introspection property. Consider the following informal argument. Suppose John is certain that the bank in open (B_J^copen). According to our interpretation of certainty, we have that John is certain that he knows that the bank is open ($B_J^c K_J$open). Suppose that the bank is in fact closed ($\neg$open). This means that John does not know that the bank is closed, because of the veridity property ($\neg K_J$open). Because of the negative introspection property, we can conclude that John knows that he does not know that the bank is open ($K_J\neg K_J$open). If we reasonably assume that knowledge implies certainty, we also get that John is certain that he does not know that the bank is open ($B_J^c\neg K_J$open). Thus we get that John is certain of both the sentence $K_J$open and its negation, which contradicts the consistency property of certainty. So something has to give—and the prime candidate is the negative introspection axiom for knowledge. And indeed, even absent a discussion of belief, this axiom has attracted criticism early on as being counter-intuitive.

This is a good point at which to introduce a caveat. Commonsense can serve at best as a crude guideline for selecting formal models. Human language and thinking are infinitely flexible, and we must accept at the outset a certain discrepancy between the meaning of a commonsense term and any formalization of it. That said, some discrepancies are more harmful than others, and negative introspection for knowledge may be one of the biggest offenders. This criticism does not diminish the importance of the partition model, which as we have seen is quite useful for modeling a variety of information settings. What it does mean is that perhaps the partition model is better thought of as capturing something other than knowledge; perhaps "possesses the implicit information that" would be a more apt descriptor of the **S5** modal operator. It also means then that we need to look for an alternative model that better captures our notion of knowledge, albeit still in a highly idealized fashion.

Here is one such model. Before we present it, one final philosophical musing. Philosophers have offered two informal slogans to explain the connection between knowledge and belief. The first is "knowledge is justified, true belief." The intuition is that knowledge is a belief that not only is true, but is held with proper justification. What "proper" means is open to debate, and we do not offer a formal account of this slogan. The second slogan is "knowledge is belief that is stable with respect to the truth." The intuition is that, when presented with evidence to the contrary, an agent would discard any of his false belief. It is only correct beliefs that cannot be contradicted by correct evidence, and are thus stable in the face of such evidence. Our formal model can be thought of a formal version of the second informal slogan; we will return to this intuition when we discuss *belief revision* later in Section 14.2.1.

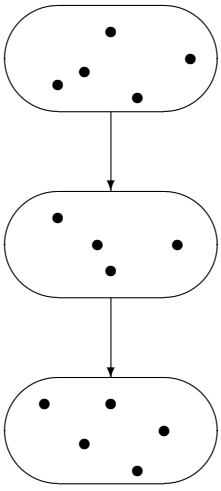

Figure 13.9 A KB structure.

In developing our combined formal model of knowledge and belief, we need to be clear about which sense of belief we intend. In the following we restrict the attention to the certainty kind, namely, B_i^c. Since there is no ambiguity, we will use the simpler notation B_i.

We have so far encountered two special classes of Kripke models—partitions and quasi-partitions. We will now introduce a third special class, the class of *total preorders* (a total preorder $\leq$ over domain Y is a reflexive and transitive binary relation, such that for all $y, y' \in Y$ it is the case that either $y \leq y'$, or $y' \leq y$, or both).

KB-structure **Definition 13.7.1 (KB- structure)** *An* (n-agent) KB-structure *is a tuple* $(W, \leq_1, \ldots, \leq_n)$, *where:*

- W *is a set of possible worlds; and*
- *each* $\leq_i$ *is a finite total preorder over* W.

KB-model *An* (n-agent) KB-model *over a language* Σ *is a tuple* $A = (W, \pi, \leq_1, \ldots, \leq_n)$, *where:*

- W *and* $\leq_i$ *are as earlier; and*
- $\pi : \Sigma \mapsto 2^W$ *is an interpretation function that determines which sentences in the languages are true in which worlds.*

Although there is merit in considering more general cases, we will confine our attention to *well-founded* total preorders; that is, we will assume that for no preorder $\leq_i$ does there exist an infinite sequence of worlds $w_1, w_2, \ldots$ such that $w_j <_i w_{j+1}$ for all $j > 0$ (where $<$ is the anti-reflexive closure of $\leq$).

A graphical illustration of a KB structure for a single agent is given in Figure 13.9. In this structure there are three clusters of pairwise connected worlds, and each world in a given cluster is connected to all worlds in lower clusters.

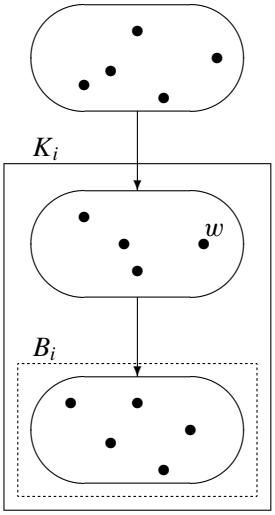

Figure 13.10 Knowledge and belief in a KB model.

We use KB structures to define both knowledge and certainty-type belief. First, it is useful to attach an intuitive interpretation to the accessibility relation. Think of it as describing the cognitive bias of an agent; $w' \leq w$ iff w' is at least as easy for the agent to imagine as w. The agent's beliefs are defined as truth in the most easily imagined worlds; intuitively, those are the only worlds the agent considers, given his current evidence. Since we are considering only finite hierarchies, these consist of the "bottom-most" cluster. However, there are other worlds that are consistent with the agent's evidence; as implausible as they are, the agent's knowledge requires truth in them as well. (Beside being relevant to defining knowledge, these worlds are relevant in the context of belief revision, discussed in Section 14.2.1.) Figure 13.10 depicts the definitions graphically; the full definitions follow.

Definition 13.7.2 (Logical entailment for KB models) *Let* $A = (W, \pi, \leq_1, \ldots, \leq_n)$ *be a KB- model over* Σ*, and* $w \in W$*. Then we define the* $\models$ *relation as:*

- *If* $\varphi \in \Sigma$*, then* $A, w \models \varphi$ *if and only if* $w \in \pi(\varphi)$*.*
- $A, w \models K_i\varphi$ *if and only if for all worlds* w'*, if* $w' \leq_i w$ *then* $A, w' \models \varphi$*.*
- $A, w \models B_i\varphi$ *if and only if* φ *is true in all worlds minimal in* $\leq_i$*.*

We can now ask whether there exists an axiom system that captures the properties of knowledge and belief, as defined by KB models. The answer is yes; the system that does it, while somewhat more complex than **S5** or **KD45**, is nonetheless illuminating. If we only wanted to capture the notion of knowledge, we would need the so-called **S4.3** axiomatic system, a well-known system of modal logic capturing total preorders. But we want to capture knowledge as well as belief. This is done by the following axioms (to which one must add the two usual inference rules of modal logic).

Knowledge:

Axiom 13.7.3 $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$

Axiom 13.7.4 $K_i\varphi \to \varphi$

Axiom 13.7.5 $K_i\varphi \to K_iK_i\varphi$

Belief:

Axiom 13.7.6 $B_i(\varphi \to \psi) \to (B_i\varphi \to B_i\psi)$

Axiom 13.7.7 $B_i\varphi \to \neg B_i\neg\varphi$

Axiom 13.7.8 $B_i\varphi \to B_iB_i\varphi$

Axiom 13.7.9 $\neg B_i\varphi \to B_i\neg B_i\varphi$

Knowledge and belief:

Axiom 13.7.10 $K_i\varphi \to B_i\varphi$

Axiom 13.7.11 $B_i\varphi \to B_iK_i\varphi$

Axiom 13.7.12 $B_i\varphi \to K_iB_i\varphi$

Axiom 13.7.13 $\neg B_i\varphi \to K_i\neg B_i\varphi$

The properties of belief are precisely the properties of a **KD45** system, the standard logic of belief. The properties of knowledge as listed consist of the **KD4** system, also known as **S4**. However, by virtue of the relationship between knowledge and belief, one obtains additional properties of knowledge. In particular, one obtains the property of the so-called **S4.2** system, a weak form of introspection:

$$\neg K_i\neg K_i\varphi \to K_i\neg K_i\neg K_i\varphi.$$

This property is unintuitive, until one notes another surprising connection that can be derived between knowledge and belief in our system:

$$\neg K_i\neg K_i\varphi \leftrightarrow B_i\varphi.$$

While one can debate the merits of this property, it is certainly less opaque than the previous one; and, if this equivalence is substituted into the previous formula, then we simply get one of the introspection properties of knowledge and belief!

13.8 History and references

The most comprehensive one-stop shop for logics of knowledge and belief is Fagin et al. [1995]. It is written by computer scientists, but covers well the perspectives of philosophy and game theory. Readers who would like to go directly to the sources should start with Hintikka [1962] in philosophy, Aumann [1976] in game theory (who introduced the partition model), Moore [1985] in computer science for an artificial intelligence perspective, and Halpern and Moses [1990] in computer science for a distributed systems perspective. The Muddy Children puzzle has its origin in the Cheating Wives puzzle in Gamow and Stern

[1958], and the Coordinated Attack problem in Gray [1978]. Another good reference is provided by the proceedings of Theoretical Aspects of Rationality and Knowledge—or TARK—which can be found online at `www.tark.org`. (The acronym originally stood for Theoretical Aspects of Reasoning about Knowledge.) There are many books on modal logic in general, not only when applied to reasoning about knowledge and belief; from the classic Chellas [1980] to more modern ones such as Fitting and Mendelsohn [1999] or Blackburn et al. [2002]. The application of logics of knowledge in robotics is based on Brafman et al. [1997] and Brafman et al. [1998].