Chapter 3

Completeness and Complexity

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

Arabian proverb

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

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

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

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

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

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

3.1 Completeness Results

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A1. All tautologies of propositional calculus

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

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

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

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

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

is a substitution instance of the Distribution Axiom.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We can now prove that K_n is sound and complete.

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

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

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

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

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

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

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

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

We proceed as follows. Given a set V of formulas, define $V/K_i = \{\varphi \mid K_i\varphi \in V\}$. For example, if $V = \{K_1p, K_2K_1q, K_1K_3p \wedge q, K_1K_3q\}$, then $V/K_1 = \{p, K_3q\}$. Let $M^c = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where

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

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

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

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

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

By R2, we have

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

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

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

Now from R1, we get

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lemma 3.1.4

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

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

(c) The following are equivalent:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proposition 3.1.6

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

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

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

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

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

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

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

3.2 Decidability

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

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

definition of decidability; references are provided in the notes at the end of the chapter.) An algorithm that recognizes valid formulas can be viewed as another characterization of the properties of knowledge, one that is complementary to the characterization in terms of a sound and complete axiom system.

A situation in which recognizing valid formulas is useful is where we have an agent whose information is characterized by a collection of formulas whose conjunction is φ. If the agent wants to know whether a formula ψ follows from the information he has, then he has to check whether $\varphi \Rightarrow \psi$ is valid. An example of this type of situation is if we take the agent to be a knowledge base. A knowledge base can draw conclusions about the state of the world based on the logical consequences of the information it has been told. (See Sections 4.4.1, 7.3, and 9.3.3 for further discussion of knowledge bases.)

A formula φ is valid in a certain class of Kripke structures if it is true in all states in all structures of that class. Thus, before examining the algorithmic aspects of validity, we consider the algorithmic aspects of truth. We refer to the problem of deciding if a formula is true in a given state of a given Kripke structure as the *model-checking problem*.

There is no general procedure for doing model checking in an infinite Kripke structure. Indeed, it is not even possible to represent arbitrary infinite structures effectively. On the other hand, in finite Kripke structures, model checking is relatively straightforward. Given a finite Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, define $||M||$, the *size* of M, to be the sum of the number of states in S and the number of pairs in $\mathcal{K}_i$, for $i = 1, \ldots, n$. In the following proposition (and in later results), we use the notation $O(f)$, read "*order* of f" or "(big) O of f" for a function f. This denotes some function g such that for each argument a, we have $g(a) \leq cf(a)$ for some constant c independent of a. Thus, for example, when we say that the running time of an algorithm is $O(||M|| \times |\varphi|)$, this means that there is some constant c, independent of the structure M and the formula φ, such that for all structures M and formulas φ the time to check if φ is satisfied in M is at most $c \times ||M|| \times |\varphi|$.

Proposition 3.2.1 *There is an algorithm that, given a structure M, a state s of M, and a formula $\varphi \in \mathcal{L}_n$, determines, in time $O(||M|| \times |\varphi|)$, whether $(M, s) \models \varphi$.*

Proof Let $\varphi_1, \ldots, \varphi_m$ be the subformulas of φ, listed in order of length, with ties broken arbitrarily. Thus we have $\varphi_m = \varphi$, and if φ_i is a subformula of φ_j, then $i < j$. There are at most $|\varphi|$ subformulas of φ (Exercise 3.1), so we must have $m \leq |\varphi|$. An easy induction on k shows that we can label each state s in M with φ_j or $\neg\varphi_j$,

for $j = 1, \ldots, k$, depending on whether or not φ_j is true at s, in time $O(k||M||)$. The only nontrivial case is if φ_{k+1} is of the form $K_i\varphi_j$, where $j < k+1$. We label a state s with $K_i\varphi_j$ iff each state t such that $(s,t) \in \mathcal{K}_i$ is labeled with φ_j. Assuming inductively that each state has already been labeled with φ_j or $\neg\varphi_j$, this step can clearly be carried out in time $O(||M||)$, as desired. ∎

Observe that this result holds independent of the number of agents. It continues to hold if we restrict attention to particular classes of structures, such as $\mathcal{M}_n^{rt}$ or $\mathcal{M}_n^{rst}$. The result can be easily extended to get a polynomial-time model-checking algorithm even if we have distributed knowledge or common knowledge in the language (Exercise 3.24). Finally, note that the algorithm can be easily modified to check whether φ holds at a particular state s in M.

It should be noted that Proposition 3.2.1 implicitly assumes a "reasonable" representation for Kripke structures. In particular, it assumes that, given a state s and a primitive proposition p, we can determine in constant time whether $\pi(s)$ assigns to p the truth value **true** or the truth value **false**. Such an assumption is not always appropriate. If s corresponds to a position in a chess game and p corresponds to "white can win from this position," then $\pi(s)(p)$ may be quite difficult to compute. Similarly, Proposition 3.2.1 requires some assumption on the time to "traverse" the edges of the Kripke structure; for example, it is sufficient to assume that given a state s where there are m edges $(s,t) \in \mathcal{K}_i$, we can find in time $O(m)$ all the states t such that $(s,t) \in \mathcal{K}_i$. These assumptions are fairly natural if we think of Kripke structures as labeled graphs, and we can read off the $\mathcal{K}_i$ relations and the states where the primitive propositions are true from the diagram describing the graph. Whenever we use a Kripke structure to model a specific situation, however, then we must reexamine these assumptions. In the case of the Kripke structure for the muddy children puzzle described in Chapter 2, it is easy to tell if a proposition p_i is true at a given state, and it is easy to compute the $\mathcal{K}_i$ relations from the descriptions of the states; in general, it may not be. We return to this issue in Chapter 10.

We now turn our attention to the problem of checking whether a given formula is provable. We start with K_n. Our first step is to show that if a formula is K_n-consistent, not only is it satisfiable in some structure (in particular, the canonical structure constructed in the proof of Theorem 3.1.3), but in fact it is also satisfiable in a finite structure (which the canonical structure is certainly not!). The proof is actually just a slight variant of the proof of Theorem 3.1.3. The idea is that rather

than considering maximal K_n-consistent subsets of $\mathcal{L}_n$ when trying to construct a structure satisfying a formula φ, we restrict attention to sets of subformulas of φ.

Theorem 3.2.2 *If $\varphi \in \mathcal{L}_n$ is K_n-consistent, then φ is satisfiable in an $\mathcal{M}_n$ structure with at most $2^{|\varphi|}$ states.*

Proof Let $Sub^+(\varphi)$ consist of all the subformulas of φ and their negations, that is, $Sub^+(\varphi) = Sub(\varphi) \cup \{\neg\psi \mid \psi \in Sub(\varphi)\}$. Let $Con(\varphi)$ be the set of maximal K_n-consistent subsets of $Sub^+(\varphi)$. A proof almost identical to that of Lemma 3.1.2 can be used to show that every K_n-consistent subset of $Sub^+(\varphi)$ can be extended to an element of $Con(\varphi)$. Moreover, a member of $Con(\varphi)$ contains either ψ or $\neg\psi$ for every formula $\psi \in Sub(\varphi)$ (but not both, for otherwise it would not be K_n-consistent). So the cardinality of $Con(\varphi)$ is at most $2^{|Sub(\varphi)|}$, which is at most $2^{|\varphi|}$, since $|Sub(\varphi)| \leq |\varphi|$.

We now construct a structure $M_\varphi = (S_\varphi, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$. The construction is essentially that of Theorem 3.1.3, except that we take $S_\varphi = \{s_V \mid V \in Con(\varphi)\}$. We can now show that if $V \in Con(\varphi)$, then for all $\psi \in Sub^+(\varphi)$ we have $(M_\varphi, s_V) \models \psi$ iff $\psi \in V$. The proof is identical to that of Theorem 3.1.3, and so is omitted here. ∎

From Theorem 3.2.2, we can get an effective (although not particularly efficient) procedure for checking if a formula φ is K_n-consistent (or equivalently, by Theorem 3.1.3, satisfiable with respect to $\mathcal{M}_n$). We simply construct every Kripke structure with $2^{|\varphi|}$ states. (The number of such structures is finite, albeit very large.) We then check if φ is true at some state of one of these structures. The latter check is done using the model-checking algorithm of Proposition 3.2.1. If φ is true at some state in one of these structures, then clearly φ is satisfiable with respect to $\mathcal{M}_n$. Conversely, if φ is satisfiable with respect to $\mathcal{M}_n$, then by Theorem 3.2.2 it must be satisfiable in one of these structures.

As a consequence, we can now show that the validity problem for $\mathcal{M}_n$ (or equivalently, by Theorem 3.1.3, the provability problem for K_n) is decidable.

Corollary 3.2.3 *The validity problem for $\mathcal{M}_n$ and the provability problem for K_n are decidable.*

Proof Since φ is provable in K_n iff $\neg\varphi$ is not K_n-consistent, we can simply check (using the aforementioned procedure) if $\neg\varphi$ is K_n-consistent. ∎

Note that by Corollary 3.2.3 we have a way of checking whether a formula is provable in K_n without deriving a single proof using the axiom system! (Actually, with some

additional effort we can extend the ideas in the proof of Theorem 3.2.2 so that if a formula is provable in K_n, then we can effectively find a proof of it; see Exercise 3.25 for details.)

We can extend the arguments of Theorem 3.2.2 to the other logics we have been considering.

Theorem 3.2.4 *If φ is T_n- (resp., $S4_n$-, $S5_n$-, $KD45_n$-) consistent, then φ is satisfiable in a structure in $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$, $\mathcal{M}_n^{elt}$) with at most $2^{|\varphi|}$ states.*

Proof The proof in the case of T_n is identical to that of Theorem 3.2.2, except that we consider maximal T_n-consistent subsets of $Sub^+(\varphi)$ rather than maximal K_n-consistent subsets of $Sub^+(\varphi)$. Note that in the case of T_n, the axiom $K_i\varphi \Rightarrow \varphi$ guarantees that $V/K_i \subseteq V$, so we get reflexivity of $\mathcal{K}_i$ even if we restrict attention to subsets of $Sub^+(\varphi)$.

The obvious modification of the proof of Theorem 3.2.2 does not work for $S4_n$, since the $\mathcal{K}_i$ relations may not be transitive if we define $(s_V, s_W) \in \mathcal{K}_i$ iff $V/K_i \subseteq W$. For example, if φ is the formula $K_1 p$, then the maximal $S4_n$-consistent subsets of $Sub^+(\varphi)$ are $V_1 = \{K_1 p, p\}$, $V_2 = \{\neg K_1 p, p\}$, and $V_3 = \{\neg K_1 p, \neg p\}$. Note that $V_1/K_1 \subseteq V_2$ and $V_2/K_1 \subseteq V_3$, but $V_1/K_1 \not\subseteq V_3$. Although $V_1/K_1 \subseteq V_2$, intuitively it should be clear that we do not want to have $(s_{V_1}, s_{V_2}) \in \mathcal{K}_1$. The reason is that every maximal $S4_n$-consistent extension of V_1 contains $K_1 K_1 p$; in such an extension, no $S4_n$-consistent extension of V_2 would be considered possible.

In the case of $S4_n$, we deal with this problem as follows: We repeat the construction of Theorem 3.2.2, except that we take $\mathcal{K}_i$ to be $\{(s_V, s_W) \mid V/K_i \subseteq W/K_i\}$. Clearly this definition guarantees that $\mathcal{K}_i$ is transitive. For $S5_n$, we take $\mathcal{K}_i$ to consist of $\{(s_V, s_W) \mid V/K_i = W/K_i\}$. This guarantees that $\mathcal{K}_i$ is an equivalence relation. In the case of $S4_n$ and $S5_n$, the axiom $K_i\varphi \Rightarrow \varphi$ guarantees that if $V/K_i \subseteq W/K_i$, then $V/K_i \subseteq W$, which we make use of in the proof. For $KD45_n$ we do not have this axiom, so we take $\mathcal{K}_i$ to consist of $\{(s_V, s_W) \mid V/K_i = W/K_i,\ V/K_i \subseteq W\}$. We leave it to the reader to check that with this definition, $\mathcal{K}_i$ is Euclidean, transitive, and serial. The proof in all cases now continues along the lines of Theorem 3.1.3; we leave details to the reader (Exercise 3.26). ∎

Just as in the case of K_n, we can use this result to give us an effective technique for deciding whether a formula is provable in T_n (resp., $S4_n$, $S5_n$, $KD45_n$).

Corollary 3.2.5 *The validity problem for $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$, $\mathcal{M}_n^{elt}$) and the provability problem for* T_n *(resp.,* $\mathrm{S4}_n$, $\mathrm{S5}_n$, $\mathrm{KD45}_n$*) are decidable.*

It turns out that in fact there are more efficient ways of checking whether a formula is provable than those suggested by the results we have just proved; we discuss this issue later in the chapter.

3.3 Incorporating Common Knowledge

We now turn our attention to axiomatizing the operators E_G and C_G. The operator C_G is "infinitary" in that it is defined as an infinite conjunction. This might suggest that we will not be able to characterize it with a finite set of axioms. Somewhat surprisingly, this turns out to be false. The axioms for common knowledge provided in Chapter 2 are complete, as we now show. This suggests that the characterization of common knowledge as a fixed point is somehow more fundamental than its characterization as an infinite conjunction. We return to this point in Chapter 11.

Let K_n^C (resp., T_n^C, $\mathrm{S4}_n^C$, $\mathrm{S5}_n^C$, $\mathrm{KD45}_n^C$) consist of all the axioms of K_n (resp., T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, $\mathrm{KD45}_n$) together with the following two axioms and inference rule taken from Chapter 2:

C1. $E_G\varphi \Leftrightarrow \bigwedge_{i\in G} K_i\varphi$

C2. $C_G\varphi \Rightarrow E_G(\varphi \wedge C_G\varphi)$

RC1. From $\varphi \Rightarrow E_G(\psi \wedge \varphi)$ infer $\varphi \Rightarrow C_G\psi$ (Induction Rule)

As the following result shows, C1, C2, and RC1 completely characterize common knowledge.

Theorem 3.3.1 *For formulas in the language $\mathcal{L}_n^C$:*

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

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

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

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

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

Proof We consider the case of K_n^C here. We can get all the other cases by modifying the proof just as we modified the proof of Theorem 3.1.3 to prove Theorem 3.1.5.

The validity of axioms C1 and C2 with respect to $\mathcal{M}_n^{rst}$, and the fact that RC1 preserves valid formulas with respect to $\mathcal{M}_n^{rst}$, was shown in Theorem 2.4.2. Although that proof was done in the context of $\mathcal{M}_n^{rst}$, as we remarked in the proof, the proof did not make use of the fact that the possibility relations were reflexive, symmetric, and transitive, and therefore it goes through without change even if we drop this assumption. So soundness follows.

For completeness, we proceed as in the proof of Theorem 3.1.3 to show that if φ is K_n^C-consistent, then φ is satisfiable in some structure in $\mathcal{M}_n$. For technical reasons that are explained below, we need to restrict to finite structures as is done in Theorem 3.2.2.

We define $Sub_C(\varphi)$ to consist of all subformulas of φ together with the formulas $E_G(\psi \wedge C_G\psi)$, $\psi \wedge C_G\psi$, and $K_i(\psi \wedge C_G\psi)$ for each subformula $C_G\psi$ of φ and $i \in G$, and the formulas $K_i\psi$ for each subformula $E_G\psi$ of φ and $i \in G$. We define $Sub_C^+(\varphi)$ to consist of all the formulas in $Sub_C(\varphi)$ and their negations, and define $Con_C(\varphi)$ to consist of all maximal K_n^C-consistent subsets of $Sub_C^+(\varphi)$. Let $M_\varphi = (S_\varphi, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where S_φ consists of $\{s_V \mid V \in Con_C(\varphi)\}$, $\pi(s_V)(p) = \mathbf{true}$ iff $p \in V$, and $\mathcal{K}_i = \{(s_V, s_W) \mid V/K_i \subseteq W\}$, $i = 1, \ldots n$. Clearly S_φ is finite; in fact, it is not hard to show that $|S_\varphi| \leq 2^{|\varphi|}$ (see Exercise 3.27).

We again want to show that for every formula $\varphi' \in Sub_C^+(\varphi)$, we have $(M_\varphi, s_V) \models \varphi'$ iff $\varphi' \in V$. We proceed by induction on the structure of formulas. The argument in the case that φ' is a primitive proposition, a conjunction, a negation, or of the form $K_i\psi$ is essentially identical to that used in Theorem 3.1.3; we do not repeat it here.

Suppose that φ' is of the form $E_G\psi$. Assume that $\varphi' \in V$. Since V is a maximal K_n^C-consistent subset of $Sub_C^+(\varphi)$, and since $Sub_C^+(\varphi)$ includes (by definition) all formulas $K_i\psi$ for $i \in G$, by C1 we get that $K_i\psi \in V$ for all $i \in G$. So by the induction hypothesis, $(M_\varphi, s_V) \models K_i\psi$ for each $i \in G$. Therefore, $(M_\varphi, s_V) \models E_G\psi$. We have shown that if $E_G\psi \in V$, then $(M_\varphi, s_V) \models E_G\psi$. The proof of the converse is very similar.

Finally, we must consider the case that φ' is of the form $C_G\psi$. That is, we need to prove that $C_G\psi \in V$ iff $(M_\varphi, s_V) \models C_G\psi$. We begin with the "only if" direction. If $C_G\psi \in V$, we show by induction on k that if s_W is G-reachable from s_V in k steps, then both ψ and $C_G\psi$ are in W. For $k = 1$, observe that axiom C2 and the fact that $V \in Con_C(\varphi)$ together imply that $E_G(\psi \wedge C_G\psi) \in V$. Now our construction

guarantees that if s_W is G-reachable from s_V in one step (so that $(s_V, s_W) \in \mathcal{K}_i$ for some $i \in G$), we have $(\psi \wedge C_G\psi) \in W$. Since $W \in Con_C(\varphi)$, it follows that both ψ and $C_G\psi$ are in W. Now assume that the claim holds for k; we prove it for $k+1$. If s_W is G-reachable from s_V in $k+1$ steps, then there exists W' such that $s_{W'}$ is G-reachable from s_V in k steps, and s_W is reachable from $s_{W'}$ in one step. By the induction hypothesis, both ψ and $C_G\psi$ are in W'. The argument for the base case now shows that both $C_G\psi$ and ψ are in W. Our argument shows that, in particular, $\psi \in W$ for all W such that s_W is G-reachable from s_V. By our main induction hypothesis, $(M_\varphi, s_W) \models \psi$ for all s_W that are G-reachable from s_V. Thus, $(M_\varphi, s_V) \models C_G\psi$.

The proof of the converse, that if $(M_\varphi, s_V) \models C_G\psi$ then $C_G\psi \in V$, is the hardest part of the proof. Assume that $(M_\varphi, s_V) \models C_G\psi$. We can describe each state s_W of M_φ by the conjunction of the formulas in W. This conjunction, which we denote by φ_W, is a formula in $\mathcal{L}_n^C$, since W is a finite set. Here we make crucial use of the fact that we restrict to formulas in $Sub_C^+(\varphi)$, a finite set, rather than consider maximal K_n^C-consistent subsets of $\mathcal{L}_n^C$, which would have been the straightforward analogue to the proof of Theorem 3.1.3. Let $\mathcal{W} = \{W \in Con_C(\varphi) \mid (M_\varphi, s_W) \models C_G\psi\}$. Define $\varphi_{\mathcal{W}}$ to be $\bigvee_{W \in \mathcal{W}} \varphi_W$. Thus, $\varphi_{\mathcal{W}}$ is the disjunction of the description of all of the states s_W where $C_G\psi$ holds, and can be thought of as the formula that characterizes these states. Since the set $\mathcal{W}$ is finite, it follows that $\varphi_{\mathcal{W}}$ is a formula in $\mathcal{L}_n^C$. The key step in the proof is to make use of the Induction Rule (RC1), where $\varphi_{\mathcal{W}}$ plays the role of φ. In Exercise 3.28, we prove that

$$\mathrm{K}_n^C \vdash \varphi_{\mathcal{W}} \Rightarrow E_G(\psi \wedge \varphi_{\mathcal{W}}). \tag{3.1}$$

Therefore, by the Induction Rule, we obtain

$$\mathrm{K}_n^C \vdash \varphi_{\mathcal{W}} \Rightarrow C_G\psi.$$

Since $V \in \mathcal{W}$, we have $\mathrm{K}_n^C \vdash \varphi_V \Rightarrow \varphi_{\mathcal{W}}$, so

$$\mathrm{K}_n^C \vdash \varphi_V \Rightarrow C_G\psi. \tag{3.2}$$

It follows that $C_G\psi \in V$, as desired. For if $C_G\psi \notin V$, then $\neg C_G\psi \in V$, and it is easy to see that this, together with (3.2), would imply that V is not K_n^C-consistent, a contradiction. ∎

Notice that our proof shows that if a formula φ is satisfiable at all, it is satisfiable in a finite structure (in fact, with at most $2^{|\varphi|}$ states.) Thus, using the techniques of

the previous section, we again get that the validity problem for K_n^C (resp., T_n^C, $\mathrm{S4}_n^C$, $\mathrm{S5}_n^C$, $\mathrm{KD45}_n^C$) is decidable.

3.4 Incorporating Distributed Knowledge

The last operator we would like to axiomatize is D_G. The major new properties of distributed knowledge were discussed in Chapter 2:

D1. $D_{\{i\}}\varphi \Leftrightarrow K_i\varphi, \quad i = 1, \ldots, n$

D2. $D_G\varphi \Rightarrow D_{G'}\varphi$ if $G \subseteq G'$

In addition, the D_G operator has all the properties of the knowledge operator. What these are depends on the system we consider. Thus, for example, in all cases A2 applies to D_G, so that the following axiom is valid:

$$(D_G\varphi \wedge D_G(\varphi \Rightarrow \psi)) \Rightarrow D_G\psi.$$

If in addition we take the $\mathcal{K}_i$ relations to be reflexive, so that knowledge satisfies A3, then so does distributed knowledge; that is, we get in addition the axiom $D_G\varphi \Rightarrow \varphi$. Similar remarks hold for A4 and A5. This, however, is not the case for A6; it is easy to see that even if the $\mathcal{K}_i$ relations are serial, their intersection may be empty. Let K_n^D (resp., T_n^D, $\mathrm{S4}_n^D$, $\mathrm{S5}_n^D$) $\mathrm{KD45}_n^D$) be the system that results from adding axioms D1, D2 to K_n (resp., T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, $\mathrm{KD45}_n$), and assuming that all of the other axioms apply to the modal operators D_G (except for A6 in the case of $\mathrm{KD45}_n^D$) as well as K_i. Thus, for example, $\mathrm{S4}_n^D$ includes the axiom $D_G\varphi \Rightarrow D_G D_G\varphi$.

Theorem 3.4.1 *For formulas in the language $\mathcal{L}_n^D$*

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

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

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

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

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

Proof The proof of soundness is straightforward (see Exercise 2.10). Although the basic ideas of the completeness proof are similar to those we have encountered before, the details are somewhat technical. We just sketch the main ideas here, leaving the details to the exercises. We consider only the case of K_n^D here.

We start with a canonical structure constructed just as in the proof of Theorem 3.1.3, treating the distributed knowledge operators D_G exactly like the K_i operators. That is, for each maximal K_n^D-consistent set V, we define the set V/D_G in the obvious way and define binary relations $\mathcal{K}_G$ on S via $(s_V, s_W) \in \mathcal{K}_G$ iff $V/D_G \subseteq W$. From axiom D1 it follows that $\mathcal{K}_{\{i\}}$ (the binary relation derived using D_G, where G is the singleton set $\{i\}$) is equal to $\mathcal{K}_i$ (the binary relation derived using K_i). It can be shown that $\mathcal{K}_G \subseteq \bigcap_{i \in G} \mathcal{K}_i$; however, in general, equality does not hold. By making multiple copies of states in the canonical structure that are in $\bigcap_{i \in G} \mathcal{K}_i$ and not in $\mathcal{K}_G$, it is possible to construct a structure at which the same formulas are true in corresponding states, and in which $\bigcap_{i \in G} \mathcal{K}_i$ and $\mathcal{K}_G$ coincide. This gives us the desired structure. (See Exercise 3.30 for further details.) ∎

We have considered axiom systems for the languages $\mathcal{L}_n^C$ and $\mathcal{L}_n^D$. It is not too hard to show that we can get sound and complete axiomatizations for the language $\mathcal{L}_n^{CD}$, which has modal operators for common knowledge and distributed knowledge, by combining the axioms for common knowledge and distributed knowledge. It can also be shown that the validity problem is decidable. There are no interesting new ideas involved in doing this, so we shall not carry out that exercise here.

3.5 The Complexity of the Validity Problem

In earlier sections, we have shown that the validity problem for the various logics we have been considering is decidable. In this section, we examine the issue more carefully. In particular, we attempt to completely characterize the inherent difficulty of deciding validity for all the logics we consider. This will give us some insight into which features of a logic make deciding validity difficult. We characterize the "inherent difficulty" of a problem in terms of computational complexity. We briefly review the necessary notions here.

Formally, we view everything in terms of the difficulty of determining membership in a set. Thus, the validity problem is viewed as the problem of determining whether a given formula φ is an element of the set of formulas valid with respect to a class of structures. The difficulty of determining set membership is usually measured

by the amount of time and/or space (memory) required to do this, as a function of the input size. Since the inputs we consider in this section are formulas, we will typically be interested in the difficulty of determining whether a formula φ is valid or satisfiable as a function of $|\varphi|$.

We are usually most interested in *deterministic* computations, where at any point in a computation, the next step of the computation is uniquely determined. However, thinking in terms of *nondeterministic* computations—ones where the program may "guess" which of a finite number of steps to take—has been very helpful in classifying the intrinsic difficulty of a number of problems. A deterministic algorithm must conclude by either accepting the input (intuitively, saying "Yes, the formula that is the input is valid") or rejecting the input (intuitively, saying "No, the formula that is the input is not valid"). A nondeterministic algorithm is said to accept an input if it accepts for some appropriate sequence of guesses.

The complexity classes we are most concerned with here are *P*, *NP*, *PSPACE*, *EXPTIME*, and *NEXPTIME*: those sets such that determining whether a given element x is a member of the set can be done in deterministic polynomial time, nondeterministic polynomial time, deterministic polynomial space, deterministic exponential time (where exponential in n means in time 2^{dn} for some constant d), and nondeterministic exponential time, respectively (as a function of the size of x). It is not hard to show that $P \subseteq NP \subseteq PSPACE \subseteq EXPTIME \subseteq NEXPTIME$. It is also known that $P \neq EXPTIME$ and that $NP \neq NEXPTIME$. While it is conjectured that all the other inclusions are strict, proving this remains elusive. The $P = NP$ problem is currently considered the most important open problem in the field of computational complexity. Interestingly, it is known that nondeterminism does not add any power at the level of polynomial space: nondeterministic polynomial space is equivalent to deterministic polynomial space.

Given a complexity class $\mathcal{C}$, the class co-$\mathcal{C}$ consists of all of the sets whose *complement* is a member of $\mathcal{C}$. Notice that if we have a deterministic algorithm $\mathbf{A}$ for deciding membership in a set A, then it is easy to convert it to an algorithm $\mathbf{A}'$ for deciding membership in the complement of A that runs in the same space and/or time bounds: $\mathbf{A}'$ accepts an input x iff $\mathbf{A}$ rejects. It follows that $\mathcal{C} =$ co-$\mathcal{C}$ must hold for every deterministic complexity class $\mathcal{C}$, in particular, for *P*, *PSPACE* and *EXPTIME*. This is not necessarily the case for a nondeterministic algorithm. There is no obvious way to construct an algorithm $\mathbf{A}'$ that will accept an element of the complement of A by an appropriate sequence of guesses. Thus, in particular, it is not

known whether $NP = \text{co-}NP$. Clearly, if $P = NP$, then it would immediately follow that $NP = \text{co-}NP$, but it is conjectured that in fact $NP \neq \text{co-}NP$.

Roughly speaking, a set A is said to be *hard* with respect to a complexity class $\mathcal{C}$ (e.g., *NP*-hard, *PSPACE*-hard, etc.) if every set in $\mathcal{C}$ can be efficiently *reduced* to A; that is, for every set B in $\mathcal{C}$, an algorithm deciding membership in B can be easily obtained from an algorithm for deciding membership in A. A set is *complete* with respect to a complexity class $\mathcal{C}$, or $\mathcal{C}$-*complete* if it is both in $\mathcal{C}$ and $\mathcal{C}$-hard.

The problem of determining whether a formula of propositional logic is satisfiable (i.e., the problem of determining whether a given propositional formula is in the set of satisfiable propositional formulas) is *NP*-complete. In particular, this means that if we could find a polynomial-time algorithm for deciding satisfiability for propositional logic, we would also have polynomial-time algorithms for all other *NP* problems. This is considered highly unlikely.

What about the complexity of determining validity? Notice that satisfiability and validity are complementary problems. A formula φ is valid exactly if $\neg\varphi$ is not satisfiable. Thus, if the satisfiability problem for a logic is complete for some complexity class $\mathcal{C}$, then the validity problem must be complete for co-$\mathcal{C}$. In particular, this means that the validity problem for propositional logic is co-*NP*-complete.

The complexity of the satisfiability and validity problem for numerous logics other than propositional logic has been studied. It is remarkable how many of these problems can be completely characterized in terms of the complexity classes discussed here. In particular, this is true for the logics we consider here. (We remark that when we speak of a *logic*, we typically mean an axiom system together with a corresponding class of structures. We usually refer to a logic by the name of the axiom system. Thus, when we speak of "the satisfiability problem for (the logic) $\mathrm{S4}_n$," we mean the problem of determining if a formula $\varphi \in \mathcal{L}_n$ is satisfiable with respect to $\mathcal{M}_n^{rt}$ or, equivalently, the problem of determining if φ is $\mathrm{S4}_n$-consistent.) The situation is summarized in Table 3.2. The results in the table are stated in terms of the satisfiability problem. Since φ is valid iff $\neg\varphi$ is not satisfiable, a solution to the validity problem quickly leads to a solution to the satisfiability problem, and vice versa. In particular, in those cases where the satisfiability problem is *PSPACE*- or *EXPTIME*-complete, the validity problem has the same complexity as the satisfiability problem. In the cases where the satisfiability problem is *NP*-complete, the validity problem is co-*NP*-complete.

As can be seen from the table, without common knowledge in the language, the satisfiability problem is in general *PSPACE*-complete. In the case of $\mathrm{S4}_2$, for

$S5_1$, $KD45_1$	K_n, T_n, $S4_n$, $n \geq 1$; $S5_n$, $KD45_n$, $n \geq 2$	K_n^C, T_n^C, $n \geq 1$; $S4_n^C$, $S5_n^C$, $KD45_n^C$, $n \geq 2$
NP-complete	*PSPACE*-complete	*EXPTIME*-complete

Table 3.2 The complexity of the satisfiability problem for logics of knowledge

example, this means that there is an algorithm for deciding whether a formula is satisfiable with respect to $\mathcal{M}_2^{rt}$ (or, equivalently, whether it is $S4_2$-consistent) that runs in polynomial space, and every *PSPACE* problem can be efficiently reduced to the satisfiability problem for $S4_2$. The only exception to the *PSPACE* result is in the case of S5 and KD45 (for only one agent), where the satisfiability problem is *NP*-complete. This says that in the case of S5 and KD45, going from one agent to many agents increases the complexity of the logic (provided that *PSPACE* $\neq$ *NP*). Adding common knowledge causes a further increase in complexity, to *EXPTIME*-complete.

We remark that we do not mention languages involving distributed knowledge in our table. This is because adding distributed knowledge to the language does not affect the complexity. Thus, for example, the complexity of the satisfiability problem for $S5_n$ is the same as that for $S5_n^D$. We also do not mention the single-agent case for $S4^C$, $S5^C$, and $KD45^C$, because in these cases common knowledge reduces to knowledge.

In the next section, we prove *NP*-completeness for S5 and KD45 in detail. The proofs for the *PSPACE* upper and lower bounds are quite technical and are beyond the scope of this book. (See the notes for references.) We remark that our lower bounds suggest that we cannot hope for automatic theorem provers for these logics that are guaranteed to work well (in the sense of providing the right answer quickly) for all inputs.

It is interesting to compare the results mentioned in the table with those proved in Section 3.2. The results there give us a nondeterministic exponential time algorithm for deciding satisfiability: given a formula φ, we simply guess an exponential-sized structure satisfying φ (if φ is satisfiable, then there must be such a structure by the results of Section 3.2) and then verify (using the model-checking algorithm) that the structure indeed satisfies φ. Since the structure is exponential-sized, the

model-checking can be done in exponential time. The algorithm is in nondeterministic exponential time because of the guess made at the beginning. Because, as we mentioned earlier, it is strongly suspected that exponential time, and hence nondeterministic exponential time, is worse than polynomial space, this suggests that the algorithm of Section 3.2 is not optimal.

3.6 *NP*-Completeness Results for S5 and KD45

In this section, we focus on the single-agent case of $S5_n$ and $KD45_n$, namely S5 and KD45. It is clear that the satisfiability problem for S5 and KD45 must be at least *NP*-hard, since it is at least as hard as the satisfiability problem for propositional logic. It is somewhat surprising that reasoning about knowledge in this case does not add any further complexity. We start with S5 here.

Theorem 3.6.1 *The satisfiability problem for S5 is NP-complete (and thus the validity problem for S5 is co-NP-complete).*

The key step in the proof of Theorem 3.6.1 lies in showing that satisfiable S5 formulas can be satisfied in structures with very few states.

Proposition 3.6.2 *An S5 formula φ is satisfiable if and only if it is satisfiable in a structure in $\mathcal{M}_1^{rst}$ with at most $|\varphi|$ states.*

Proof Suppose that $(M, s) \models \varphi$. By Proposition 3.1.6, we can assume without loss of generality that $M = (S, \pi, \mathcal{K}_1)$, where $\mathcal{K}_1$ is a universal relation, so that $(t, t') \in \mathcal{K}_1$ for all $t, t' \in S$. Let F be the set of subformulas of φ of the form $K_1\psi$ for which $(M, s) \models \neg K_1\psi$; that is, F is the set of subformulas of φ that have the form $K_1\psi$ and are false at the state s. For each formula $K_1\psi \in F$, there must be a state $s_\psi \in S$ such that $(M, s_\psi) \models \neg\psi$. Let $M' = (S', \pi', \mathcal{K}_1')$, where (a) $S' = \{s\} \cup \{s_\psi \mid \psi \in F\}$, (b) π' is the restriction of π to S', and (c) $\mathcal{K}_1' = \{(t, t') \mid t, t' \in S'\}$. Since $|F| < |Sub(\varphi)| \leq |\varphi|$, it follows that $|S'| \leq |\varphi|$. We now show that for all states $s' \in S'$ and for all subformulas ψ of φ (including φ itself), $(M, s') \models \psi$ iff $(M', s') \models \psi$. As usual, we proceed by induction on the structure of ψ. The only nontrivial case is when ψ is of the form $K_1\psi'$. Suppose that $s' \in S'$. If $(M, s') \models K_1\psi'$, then $(M, t) \models \psi'$ for all $t \in S$, so, in particular, $(M, t) \models \psi'$ for all $t \in S'$. By the induction hypothesis, $(M', t) \models \psi'$ for all

$t \in S'$, so $(M', s') \models K_1\psi'$. For the converse, suppose that $(M, s') \not\models K_1\psi'$. Then $(M, t) \models \neg\psi'$ for some $t \in S$. Because $\mathcal{K}_1$ is the universal relation, it follows that $(s, t) \in \mathcal{K}_1$, so $(M, s) \models \neg K_1\psi'$. But then it follows that $K_1\psi' \in F$, and $(M, s_{\psi'}) \models \neg\psi'$. By construction, $s_{\psi'} \in S'$, and by the induction hypothesis, we also have $(M', s_{\psi'}) \models \neg\psi'$. Because $(s', s_{\psi'}) \in \mathcal{K}'_1$, we have $(M', s') \models \neg K_1\psi'$, and so $(M', s') \not\models K_1\psi'$ as desired. This concludes the proof that $(M, s') \models \psi$ iff $(M', s') \models \psi$ for all subformulas ψ of φ and all $s' \in S'$. Since $s \in S'$ and $(M, s) \models \varphi$ by assumption, we also have $(M', s) \models \varphi$. ∎

Proof of Theorem 3.6.1 As we have already mentioned, S5 satisfiability is clearly *NP*-hard. We now give an *NP* algorithm for deciding S5 satisfiability. Intuitively, given a formula φ, we simply guess a structure $M \in \mathcal{M}_n^{rst}$ with a universal possibility relation and at most $|\varphi|$ states, and verify that φ is satisfied in M. More formally, we proceed as follows. Given a formula φ, where $|\varphi| = m$, we nondeterministically guess a structure $M = (S, \pi, \mathcal{K}_1)$, where S is a set of $k \leq m$ states, $(s, t) \in \mathcal{K}_1$ for all $s, t \in S$, and for all $s \in S$ and all primitive propositions p not appearing in φ, we have $\pi(s)(p) = \mathbf{false}$. (Note that the only "guessing" that enters here is the choice of k and the truth values $\pi(s)(q)$ that the primitive propositions q appearing in φ have in the k states of S.) Since at most m primitive propositions appear in φ, guessing such a structure can be done in nondeterministic time $O(m^2)$ (i.e., at most cm^2 steps for some constant c). Next, we check whether φ is satisfied at some state $s \in S$. By Proposition 3.2.1, this can be done deterministically in time $O(m^3)$. By Proposition 3.6.2, if φ is satisfiable, it must be satisfiable in one of the structures of the kind we guessed. (Of course, if φ is not satisfiable, no guess will be right.) Thus, we have a nondeterministic $O(m^3)$ algorithm for deciding if φ is satisfiable. ∎

We can prove essentially the same results for KD45 as for S5. Using Proposition 3.1.6, we can show the following:

Proposition 3.6.3 *A KD45 formula φ is satisfiable iff it is satisfiable in a structure in $\mathcal{M}_1^{elt}$ with at most $|\varphi|$ states.*

Proof See Exercise 3.34. ∎

Using Proposition 3.6.3 just as we used Proposition 3.2.1, we can now prove the following theorem:

Theorem 3.6.4 *The satisfiability problem for KD45 is NP-complete (and thus the validity problem for KD45 is co-NP-complete).*

Proof See Exercise 3.34. ∎

We remark that results similar to Proposition 3.6.3 and Theorem 3.6.4 can also be proved for K45 (Exercise 3.35).

3.7 The First-Order Logic of Knowledge

So far, we have considered only *propositional* modal logic. That is, the formulas we have allowed contain only primitive propositions, together with propositional connectives such as $\wedge$ and $\neg$, and modal operators such as K_i and C. *First-order logic* has greater expressive power than propositional logic. It allows us to reason about individuals in a domain and properties that they have. Among other things, first-order logic allows us to express properties that all individuals have and that some individuals have, by using a universal quantifier ($\forall$, or "for all") and an existential quantifier ($\exists$, or "there exists"). For example, we can say that Pete is the governor of California using a formula such as $Governor(California, Pete)$. To say that every state has a governor, we might write the first-order formula $\forall x(State(x) \Rightarrow \exists y Governor(x, y))$: for all states x, there exists y such that the governor of x is y. First-order logic is, in a precise sense, expressive enough to capture all of propositional modal logic (see Exercise 3.37). By combining the quantifiers of first-order logic and the modal operators of propositional modal logic, we get a yet more expressive logic, *first-order modal logic*. For example, neither in first-order logic nor in propositional modal logic can we say that Alice knows that California has a governor. We can, however, say this in first-order modal logic with the formula

$$K_{Alice}\exists y Governor(California, y).$$

There are some subtleties involved in combining first-order quantifiers with modal operators. We briefly discuss them in this section, to give the reader a feeling for the issues that arise.

Despite its additional power, we make relatively little use of first-order modal logic in the rest of the book, both because most of the examples that we discuss can be expressed using propositional modal logic and because most of the issues that we are interested in already arise in the propositional case. Nevertheless, the first-order case may well be important for more complex applications.

3.7.1 First-Order Logic

In this section we briefly review first-order logic. The reader familiar with first-order logic may still want to skim this section to get acquainted with our notation.

In propositional logic, we start with a set Φ of primitive propositions. In first-order logic, we start with a set $\mathcal{T}$ of *relation symbols*, *function symbols*, and *constant symbols*. Each relation symbol and function symbol has some *arity*, which intuitively corresponds to the number of arguments it takes. If the arity is k, then we refer to the symbol as k-ary. In our earlier example, the relation symbol *Governor* has arity 2: that is, $Governor(x, y)$ has two arguments, x and y. A function symbol *Gov*, where intuitively $Gov(x) = y$ means that the governor of state x is person y, has arity 1, since it takes only one argument, namely x. We refer to the set of relation symbols, function symbols, and constant symbols as the *vocabulary*.

We assume an infinite supply of *variables*, which we usually write as x and y, possibly along with subscripts. Constant symbols and variables are both used to denote individuals in the domain. We can also form more complicated terms denoting individuals by using function symbols. Formally, the set of *terms* is formed by starting with variables and constant symbols, and closing off under function application, so that if f is a k-ary function symbol, and if $t_1, \ldots, t_k$ are terms, then $f(t_1, \ldots, t_k)$ is a term. We need terms to define formulas. An *atomic formula* is either of the form $P(t_1, \ldots, t_k)$, where P is a k-ary relation symbol and $t_1, \ldots, t_k$ are terms, or of the form $t_1 = t_2$, where t_1 and t_2 are terms. As in propositional logic, if φ and ψ are formulas, then so are $\neg\varphi$ and $\varphi \wedge \psi$. In addition, we can form new formulas using quantifiers. If φ is a formula and x is a variable, then $\exists x\varphi$ is also a formula. We use the same abbreviations as we did in the propositional case. For example, $\varphi_1 \vee \varphi_2$ is an abbreviation for $\neg(\neg\varphi_1 \wedge \neg\varphi_2)$. Furthermore, we write $\forall x\varphi$ as an abbreviation for $\neg\exists x\neg\varphi$.

We give semantics to first-order formulas using *relational structures*. Roughly speaking, a relational structure consists of a domain of individuals and a way of associating with each of the elements of the vocabulary corresponding entities over the domain. Thus, a constant symbol is associated with an element of the domain, a function symbol is associated with a function on the domain, and so on. More precisely, fix a vocabulary $\mathcal{T}$. A *relational $\mathcal{T}$-structure* (which we sometimes simply call a relational structure or just a structure) $\mathcal{A}$ consists of a nonempty set, denoted $dom(\mathcal{A})$, called the *domain*, an assignment of a k-ary relation $P^{\mathcal{A}} \subseteq dom(\mathcal{A})^k$ to each k-ary relation symbol P of $\mathcal{T}$, an assignment of a k-ary function $f^{\mathcal{A}} : dom(\mathcal{A})^k \rightarrow dom(\mathcal{A})$

to each k-ary function symbol f of $\mathcal{T}$, and an assignment of a member $c^{\mathcal{A}}$ of the domain to each constant symbol c. We refer to $P^{\mathcal{A}}$ as the *interpretation* of P in $\mathcal{A}$, and similarly for function symbols and constant symbols.

As an example, suppose that $\mathcal{T}$ consists of one binary relation symbol E. In that case, a $\mathcal{T}$-structure is simply a graph. (Recall that a graph consists of a set of nodes, some of which are connected by edges.) The domain is the set of nodes of the graph, and the interpretation of E is the edge relation of the graph, so that there is an edge from a_1 to a_2 exactly if $(a_1, a_2) \in E^{\mathcal{A}}$.

Notice that a relational structure does not provide an interpretation of the variables. Technically, it turns out to be convenient to have a separate function that does this. A *valuation* V on a structure $\mathcal{A}$ is a function from variables to elements of $dom(\mathcal{A})$. Recall that we suggested that terms are intended to represent members of the domain of a structure $\mathcal{A}$. Given a structure $\mathcal{A}$ and a valuation V on $\mathcal{A}$, we can inductively extend V in a straightforward way to a function, denoted $V^{\mathcal{A}}$ (although we frequently omit the superscript $\mathcal{A}$ when it is clear from context), that maps terms to elements of $dom(\mathcal{A})$, as follows. Define $V^{\mathcal{A}}(c) = c^{\mathcal{A}}$ for each constant symbol c, and then extend the definition by induction on the structure of terms, by taking $V^{\mathcal{A}}(f(t_1, \ldots, t_k)) = f^{\mathcal{A}}(V^{\mathcal{A}}(t_1), \ldots, V^{\mathcal{A}}(t_k))$.

With these definitions in hand, we can now define what it means for a formula to be true in a relational structure. Before we give the formal definition, we consider a few examples. Let *Tall* be a unary relation symbol, and let *President* be a constant symbol. What does it mean for $Tall(President)$ to be true in the structure $\mathcal{A}$? If we think of the domain of $\mathcal{A}$ as consisting of people, then the interpretation $Tall^{\mathcal{A}}$ of the *Tall* relation can be thought of intuitively as the set of all tall people in the domain. Assume that $President^{\mathcal{A}} = Bill$, so that, intuitively, the president is Bill. Then $Tall(President)$ is true in the structure $\mathcal{A}$ precisely if *Bill* is in the relation $Tall^{\mathcal{A}}$, that is, intuitively, if Bill is tall. How should we deal with quantification? In particular, what should it mean for $\exists x\, Tall(x)$ to be true in the structure $\mathcal{A}$? Intuitively, this formula is true when there exists a tall person in the domain of $\mathcal{A}$. Formally, $\exists x\, Tall(x)$ is true in the structure $\mathcal{A}$ precisely if the relation $Tall^{\mathcal{A}}$ is nonempty. Similarly, $\forall x\, Tall(x)$ is true in the structure $\mathcal{A}$ precisely if the relation $Tall^{\mathcal{A}}$ contains every member of the domain of $\mathcal{A}$, that is, if everyone is tall. As a final example, consider the formula $Governor(c, x)$, where c is a constant symbol and x is a variable. It does not make sense to ask whether or not the formula $Governor(c, x)$ is true in a structure $\mathcal{A}$ without knowing what value x takes on. Here we make use of valuations, which assign to each variable a member of the domain of the structure. Thus, rather than ask whether $Governor(c, x)$ is

true in a structure $\mathcal{A}$, we instead ask whether $Governor(c, x)$ is true in a structure $\mathcal{A}$ *under a given valuation* V. Assume that $c^{\mathcal{A}} = California$ and $V(x) = Pete$, so that c takes the value *California* in the structure $\mathcal{A}$ and x takes the value *Pete* under V. Then we say that $Governor(c, x)$ is true in the structure $\mathcal{A}$ under the valuation V precisely if $(V(c), V(x)) = (c^{\mathcal{A}}, Pete) = (California, Pete) \in Governor^{\mathcal{A}}$. Intuitively, $Governor(c, x)$ is true in the structure $\mathcal{A}$ under the valuation V iff Pete is the governor of California according to the structure $\mathcal{A}$.

We now give the formal semantics. Fix a relational structure $\mathcal{A}$. For each valuation V on $\mathcal{A}$ and formula φ, we define what it means for φ to be true in $\mathcal{A}$ under the valuation V, written $(\mathcal{A}, V) \models \varphi$. If V is a valuation, x is a variable, and $a \in dom(\mathcal{A})$, then let $V[x/a]$ be the valuation V' such that $V'(y) = V(y)$ for every variable y except x, and $V'(x) = a$. Thus, $V[x/a]$ agrees with V, except possibly on x, and it assigns the value a to x.

- $(\mathcal{A}, V) \models P(t_1, \ldots, t_k)$, where P is a k-ary relation symbol and $t_1, \ldots, t_k$ are terms, iff $(V(t_1), \ldots, V(t_k)) \in P^{\mathcal{A}}$
- $(\mathcal{A}, V) \models (t_1 = t_2)$, where t_1 and t_2 are terms, iff $V(t_1) = V(t_2)$
- $(\mathcal{A}, V) \models \neg\varphi$ iff $(\mathcal{A}, V) \not\models \varphi$
- $(\mathcal{A}, V) \models \varphi_1 \wedge \varphi_2$ iff $(\mathcal{A}, V) \models \varphi_1$ and $(\mathcal{A}, V) \models \varphi_2$
- $(\mathcal{A}, V) \models \exists x\varphi$ iff $(\mathcal{A}, V[x/a]) \models \varphi$ for some $a \in dom(\mathcal{A})$.

Recall that $\forall x\varphi$ is taken to be an abbreviation for $\neg\exists x\neg\varphi$. It is easy to see that $(\mathcal{A}, V) \models \forall x\varphi$ iff $(\mathcal{A}, V[x/a]) \models \varphi$ for every $a \in dom(\mathcal{A})$ (Exercise 3.38).

To decide whether the formula $Tall(President)$ is true in the structure $\mathcal{A}$ under the valuation V, the role of V is irrelevant. That is, $(\mathcal{A}, V) \models Tall(President)$ iff $(\mathcal{A}, V') \models Tall(President)$, where V and V' are arbitrary valuations. A similar comment applies to the formula $\exists x Tall(x)$. However, this is not the case for the formula $Governor(c, x)$, where c is a constant symbol and x is a variable. There may be valuations V and V' such that $(\mathcal{A}, V) \models Governor(c, x)$ but $(\mathcal{A}, V') \not\models Governor(c, x)$, so that $V(x)$ is the governor of California, but $V'(x)$ is not.

To understand better what is going on here, we need some definitions. Roughly speaking, we say that an occurrence of a variable x in φ is *bound* by the quantifier $\forall x$ in a formula such as $\forall x\varphi$ or by $\exists x$ in $\exists x\varphi$, and that an occurrence of a variable in a formula is *free* if it is not bound. (A formal definition of what it means for an

occurrence of a variable to be free is given in Exercise 3.39.) A formula in which no occurrences of variables are free is called a *sentence*. Observe that x is free in the formula $Governor(c, x)$, but no variables are free in the the formulas $Tall(President)$ and $\exists x Tall(x)$, so the latter two formulas are sentences. It is not hard to show that the valuation does not affect the truth of a sentence. That is, if φ is a sentence, and V and V' are valuations on the structure $\mathcal{A}$, then $(\mathcal{A}, V) \models \varphi$ iff $(\mathcal{A}, V') \models \varphi$ (Exercise 3.39). In other words, a sentence is true or false in a structure, independent of the valuation used.

3.7.2 First-Order Modal Logic

Just as the syntax of propositional modal logic is obtained from that of propositional logic by adding the modal operators K_i, we get the syntax of first-order modal logic from that of first-order logic by adding the modal operators K_i. Thus, we define the language of first-order modal logic by taking the definition we gave for first-order formulas and also closing off under the modal operators $K_1, \ldots, K_n$, so that if φ is a first-order modal formula, then so is $K_i\varphi$. For example, $\forall x(K_1 Red(x))$ is a first-order modal formula, which intuitively says that for every x, agent 1 knows that x is red.

The semantics of first-order modal logic uses *relational Kripke structures*. In a (propositional) Kripke structure, each world is associated with a truth assignment to the primitive propositions via the function π. In a relational Kripke structure, the π function associates with each world a relational structure. Formally, we define a relational Kripke structure for n agents over a vocabulary $\mathcal{T}$ to be a tuple $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where S is a set of *states*, π associates with each state in S a $\mathcal{T}$-structure (i.e., $\pi(s)$ is a $\mathcal{T}$-structure for each state $s \in S$), and $\mathcal{K}_i$ is a binary relation on S.

The semantics of first-order modal logic is, for the most part, the result of combining the semantics of first-order logic and the semantics of modal logic in a straightforward way. But there are a few subtleties, as we shall see. We begin with some examples. Just as in the propositional case, we would like a formula $K_i\varphi$ to be true at a state s of a relational Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ precisely if φ is true at every state t such that $(s, t) \in \mathcal{K}_i$. As an example, let φ be the formula $Tall(President)$. In some states t of the relational Kripke structure the president might be Bill (that is, $President^{\pi(t)} = Bill$), and in some states t the president might be George. We would like the formula $K_i Tall(President)$ to be true if the president is

tall in every world that agent i considers possible, even if the president is a different person in different worlds. It is quite possible for agent i to know that the president is tall without knowing exactly who the president is.

What about the formula $K_i Tall(x)$, where x is a variable? Since x is a variable, we need a valuation to decide whether this formula is true at a given state. Assume that $V(x) = Bill$. For $K_i Tall(x)$ to be true, we want Bill to be tall in every world that agent i considers possible. But now there is a problem: although *Bill* may be a member of the domain of the relational structure $\pi(s)$, it is possible that *Bill* is not a member of the domain of $\pi(t)$ for some state t that agent i considers possible at state s. To get around this problem, we restrict attention for now to *common-domain Kripke structures*, that is, relational Kripke structures where the domain is the same at every state. We discuss the implications of this restriction in more detail later.

Under the restriction to common-domain structures, defining truth of formulas becomes quite straightforward. Fix a relational Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, where the states have common domain U. A *valuation* V on M is a function that assigns to each variable a member of U. This means that $V(x)$ is independent of the state, although the interpretation of, say, a constant c does depend on the state. As we shall see, this lets us identify the same individual in the domain at different states and plays an important role in the expressive power of first-order modal logic. For a state s of M, a valuation V on M, and a formula φ, we define what it means for φ to be true at the state s of M under the valuation V, written $(M, s, V) \models \varphi$. Most of the definitions are just as in the first-order case, where the role of $\mathcal{A}$ is played by $\pi(s)$. For example,

$(M, s, V) \models P(t_1, \ldots, t_k)$, where P is a k-ary relation symbol and $t_1, \ldots, t_k$ are terms, iff $(V^{\pi(s)}(t_1), \ldots, V^{\pi(s)}(t_k)) \in P^{\pi(s)}$.

In the case of formulas $K_i\varphi$, the definition is just as in the propositional case in Chapter 2:

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

As we said earlier, first-order modal logic is significantly more expressive than either first-order logic or propositional modal logic. One important example of its extra expressive power is that it allows us to distinguish between "knowing that" and "knowing who." For example, the formula $K_{Alice}\exists x(x = President)$ says that Alice knows that someone is the president. This formula is valid (since $\exists x(x = President)$ is valid). In particular, the formula is true even in a world where Alice does not

know whether Bill or George is the president; she may consider one world possible where Bill is the president, and consider another world possible where George is the president. Thus, although Alice knows that there is a president, she may not know exactly who the president is. To say that Alice knows who the president is, we would write $\exists x\, K_{Alice}(x = President)$. Because a valuation is independent of the state, it is easy to see that this formula says that there is one particular person who is president in every world that Alice considers possible. Notice that the fact that the valuation is independent of the state is crucial in allowing us to distinguish "knowing that" from "knowing who."

3.7.3 Assumptions on Domains

We restricted attention in the previous section to common-domain Kripke structures. This means that although we allow the interpretations of relation symbols, of function symbols, and of constant symbols to vary from state to state in a given relational Kripke structure, we do not allow the domains to vary. Essentially, this assumption says that the domain is common knowledge. This assumption is quite reasonable in many applications. When analyzing a card game, players typically have common knowledge about which cards are in the deck. Nevertheless, there are certainly applications where the domain is not common knowledge. For example, although there are no unicorns in this world, we might like to imagine possible worlds where unicorns exist. On a more practical level, if our agent is a knowledge base reasoning about the employees in a company, then the agent may not know exactly how many employees the company has.

As we saw earlier, this assumption of a common domain arose in response to a technical problem, that of making sense of the truth value of a formula where a free variable appears in the scope of a modal operator, such as in the formula $K_i Tall(x)$. Without the common-domain assumption, to decide if $K_i Tall(x)$ is true at a state s under a valuation V where $V(x) = Bill$, we have to consider the truth of $Tall(x)$ at a state t where *Bill* may not be in the domain. It is not clear what the truth value of $K_i Tall(x)$ should be in this case.

We can solve this problem by making a somewhat weaker assumption than the common-domain assumption. It suffices to assume that if world t is considered possible in world s, then the domain corresponding to s is a subset of the domain corresponding to t. Formally, this assumption says that if $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ is a relational Kripke structure and $(s, t) \in \mathcal{K}_i$, then $dom(\pi(s)) \subseteq dom(\pi(t))$.

Informally, this assumption says that every object that exists in a world s also exists in every world considered possible at s. We call this the *domain-inclusion assumption*.

While the domain-inclusion assumption lets us deal with more cases than the common-domain assumption, and does avoid the technical problems discussed above, it certainly does not handle all problems. For one thing, an agent cannot consider possible a world with fewer domain elements. This means that if we take the $\mathcal{K}_i$'s to be equivalence relations, as we have claimed that we want to do for many applications, or even just Euclidean relations, then the domain-inclusion assumption implies that in all worlds considered possible from a given world the domains must be the same. Thus, with the additional assumption that the relation is Euclidean, we cannot model in this framework the examples that we mentioned earlier involving unicorns or employees in a company.

Many solutions have been proposed for how to give a semantics without any assumptions whatsoever about relationships between domains of worlds within a relational Kripke structure. Nevertheless, it is fair to say that no solution has been universally accepted. Each proposed solution suffers from various problems. One proposed solution and a problem from which it suffers are discussed in Exercise 3.40.

3.7.4 Properties of Knowledge in Relational Kripke Structures

We now consider the properties of knowledge in relational Kripke structures. Just as before, we do this in terms of the formulas that are valid in relational Kripke structures. In the first-order case, we say that a formula is valid if it is true at every state of every relational Kripke structure under every valuation. To simplify the discussion, we assume a common domain.

In the propositional case, we saw that a sound and complete axiomatization could be obtained by considering all tautologies of propositional logic, together with some specific axioms about knowledge. It is easy to see that all the axioms of K_n are still valid in relational Kripke structures (Exercise 3.41). It is also intuitively clear that these axioms are not complete. We clearly need some axioms for first-order reasoning.

We might hope that we can get a complete axiomatization by adding all (substitution instances of) valid first-order formulas. Unfortunately, this results in an unsound system. There are two specific axioms of first-order logic that cause problems.

To state them we need a little notation. Suppose that $\varphi(x)$ is a first-order formula in which some occurrences of x are free. Let t, t_1, and t_2 be terms, and let $\varphi(t)$ be

the result of substituting t for all free occurrences of x. Assume for simplicity that no variable occurring in t, t_1, or t_2 is quantified in φ (so that, for example, for every variable y in t there is no subformula of φ of the form $\exists y\psi$; without this assumption, we may have inadvertent binding of the y in t by $\exists y$). Consider the following two axioms:

$$\varphi(t) \Rightarrow \exists x \varphi(x) \tag{3.3}$$

$$(t_1 = t_2) \Rightarrow (\varphi(t_1) \Leftrightarrow \varphi(t_2)) \tag{3.4}$$

It is easy to see that both of these axioms are valid in relational structures (Exercise 3.42). For the first one, if $\varphi(t)$ is true, then there is certainly some value we can assign to x that makes $\varphi(x)$ true, namely, the interpretation of t. Axiom (3.4) just says that "equals can be replaced by equals." As an example, taking $\varphi(x)$ to be $\mathit{Governor}(\mathit{California}, x)$, we have that $((x_1 = x_2) \Rightarrow (\mathit{Governor}(\mathit{California}, x_1)) \Leftrightarrow \mathit{Governor}(\mathit{California}, x_2))$ is valid. Although these axioms are valid in relational Kripke structures if $\varphi(x)$ is a first-order formula, we now show that neither axiom is valid if we allow φ to be an arbitrary modal formula.

We start with the first axiom. Let $\varphi(x)$ be the modal formula $K_i(\mathit{Tall}(x))$ and let t be *President*. With this substitution, the axiom becomes

$$K_i(\mathit{Tall}(\mathit{President})) \Rightarrow \exists x K_i(\mathit{Tall}(x)). \tag{3.5}$$

As we noted earlier, the left-hand side of (3.5) is true if, intuitively, the president is tall in every world that agent i considers possible, even if the president is a different person in different worlds. The right-hand side of (3.5) is, however, false if there is no one person who is tall in every possible world. Since it is possible simultaneously for the left-hand side of (3.5) to be true and the right-hand side to be false, it follows that (3.5) is not valid.

What is going on is that the valuation is independent of the state, and hence under a given valuation, a variable x is a *rigid designator*, that is, it denotes the same domain element in every state. On the other hand, a constant symbol such as *President* can denote different domain elements in distinct states. It is easy to see that (3.3) is valid if we restrict the term t to being a variable. More generally, we can show that (3.3) is valid if t is a rigid designator (Exercise 3.42).

To see that the second axiom is not valid in relational Kripke structures, let φ be $K_i(t_1 = x)$. Then the axiom becomes

$$(t_1 = t_2) \Rightarrow (K_i(t_1 = t_1) \Leftrightarrow K_i(t_1 = t_2)).$$

It is easy to see that $K_i(t_1 = t_1)$ is valid, so the axiom reduces to

$$(t_1 = t_2) \Rightarrow K_i(t_1 = t_2). \tag{3.6}$$

There is a famous example from the philosophical literature that shows that this is not valid. Because of its brightness, the planet Venus is called the morning star (at sunrise, when it appears in the east), and it is also called the evening star (at sunset, when it appears in the west). Ancient astronomers referred to the morning star as Phosphorus, and the evening star as Hesperus, and were unaware that Phosphorus and Hesperus were one and the same. Let the constant symbol *Phosphorus* play the role of t_1 in (3.6), let the constant symbol *Hesperus* play the role of t_2, and let agent i be an ancient astronomer. Then (3.6) is falsified: although Hesperus and Phosphorus are equal in the real world, the astronomer does not know this.

Notice that, again, the problem here arises because t_1 and t_2 may not be rigid designators. If we restrict attention to terms that are rigid designators, and, in particular, to variables, then (3.4) is valid in all relational Kripke structures (Exercise 3.42). It follows that the following special case of (3.6), called *Knowledge of Equality*, is valid:

$$(x_1 = x_2) \Rightarrow K_i(x_1 = x_2). \tag{3.7}$$

We remark that (3.3) and (3.4) are the only axioms of first-order logic that are not valid in relational Kripke structures. More precisely, there is a complete axiomatization of first-order logic that includes (3.3) and (3.4) as axioms such that all substitution instances of all axioms besides (3.3) and (3.4) are valid in relational Kripke structures.

Suppose that we restrict (3.3) and (3.4) so that if φ is a modal formula (that is, it has occurrences of K_i operators), then the terms t, t_1, and t_2 must be variables; we henceforth call these the *restricted versions* of (3.3) and (3.4). Note that the restricted versions of (3.3) and (3.4) are valid in relational Kripke structures. We might hope that by taking (substitution instances of) the axioms of first-order logic, using only the restricted versions of (3.3) and (3.4), together with the axioms and inference rules of K_n, we would have a sound and complete axiomatization for knowledge in first-order relational structures. The resulting system is sound, but it is not complete; there are two additional axioms we must add.

One new axiom arises because of the interaction between the first-order quantifier $\forall$ and the modal operator K_i, which can be thought of as a "knowledge quantifier."

Consider the following formula, sometimes called the *Barcan formula*:

$$\forall x_1 \ldots \forall x_k K_i \varphi \Rightarrow K_i \forall x_1 \ldots \forall x_k \varphi.$$

It is fairly easy to see that the Barcan formula is valid (Exercise 3.43). Its validity, however, depends crucially on the common-domain assumption. For example, consider a relational Kripke structure whose common domain consists of precisely three elements, a_1, a_2, and a_3. Assume that Alice knows that a_1 is red, that a_2 is red, and that a_3 is red. Then, for all x, Alice knows that x is red; that is, $\forall x(K_A Red(x))$ holds. From the Barcan formula it follows that Alice knows that for every x, x is red; that is, $K_A(\forall x Red(x))$ holds. Without the common-domain assumption, we might argue intuitively that Alice does not know that every object is red, since Alice might consider it possible that there is a fourth object a_4 that is blue. In the presence of the common-domain assumption, Alice knows that a_1, a_2, and a_3 are the only domain elements, so this argument cannot be applied. On the other hand, the Barcan formula is not valid under the domain-inclusion assumption that we discussed earlier, where there really can be a fourth (non-red) object a_4 in another world (Exercise 3.44).

The second new axiom arises because of the interaction between the K_i operator and equality. This axiom, which is analogous to Knowledge of Equality (3.7), is called *Knowledge of Inequality*:

$$(x_1 \neq x_2) \Rightarrow K_i(x_1 \neq x_2). \tag{3.8}$$

Like Knowledge of Equality, this axiom is valid (Exercise 3.45). Unlike Knowledge of Equality, this axiom does not follow from the other axioms.

It turns out that no further new axioms beyond the Barcan formula and Knowledge of Inequality are needed to get a sound and complete axiomatization for the first-order theory of knowledge. Such an axiomatization (for structures with n agents) is obtained by combining

(a) the axiom system K_n,

(b) the axiom system for first-order logic referred to previously, except that we use the restricted versions of (3.3) and (3.4),

(c) the Barcan formula, and

(d) Knowledge of Inequality.

Notice that if we do not allow function or constant symbols in the vocabulary, then the only terms are variables, which are rigid designators. In this case, all substitution instances of axioms (3.3) and (3.4) are valid (in fact, the restricted versions of (3.3) and (3.4) are identical to the unrestricted versions), so we can simplify the statement of (b) above.

We have already seen that for Kripke structures $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, additional properties of the $\mathcal{K}_i$ relations give us additional axioms for knowledge. Not surprisingly, the same is true for relational Kripke structures. For example, if each $\mathcal{K}_i$ is an equivalence relation, then we can modify the sound and complete axiomatizations we just described by replacing the axiom system K_n by the axiom system $\mathrm{S5}_n$. It is interesting that in this case we do not need to include the Barcan formula or the Knowledge of Inequality axiom, since they turn out to be consequences of the axioms of first-order logic, along with $\mathrm{S5}_n$ (see Exercises 3.43 and 3.45).

Exercises

3.1 Show that $|Sub(\varphi)| \leq |\varphi|$.

3.2 Show that if neither $F \cup \{\varphi\}$ nor $F \cup \{\neg\varphi\}$ is AX-consistent, then neither is $F \cup \{\varphi \vee \neg\varphi\}$.

3.3 Prove that a maximal AX-consistent set has properties (c) and (d), as claimed in the statement of Lemma 3.1.2.

3.4 Prove that K_n is sound for $\mathcal{M}_n$, using Theorem 3.1.1.

3.5 In the proof of Theorem 3.1.3, prove that $(M^c, s_V) \models \varphi$ iff $\varphi \in V$, in the case that φ is a conjunction or a negation.

3.6 Show that if $\{\varphi_1, \ldots, \varphi_k, \neg\psi\}$ is not K_n-consistent, then

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

3.7 Prove, using induction on k together with axiom A2 and propositional reasoning, that

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

*** 3.8** Let K_n' be the variant of K_n consisting of one inference rule, modus ponens (R1), and the following two axioms:

A1′. $K_{i_1} \ldots K_{i_k}\varphi$, where φ is an instance of a tautology of propositional calculus, $k \geq 0$, and $i_1, \ldots, i_k$ are arbitrary (not necessarily distinct) agents in $\{1, \ldots, n\}$,

A2′. $K_{i_1} \ldots K_{i_k}[(K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi], i = 1, \ldots, n$, where, again, $k \geq 0$ and $i_1, \ldots, i_k$ are arbitrary (not necessarily distinct) agents in $\{1, \ldots, n\}$.

Thus, A1′ and A2′ look just like A1 and A2, except that a string of knowledge operators has been appended to the beginning of each formula. If we take $k = 0$ in each of A1′ and A2′, we get back A1 and A2.

Show that K_n is equivalent to K_n'; that is, show that a formula φ is provable in K_n iff it is provable in K_n'. Then show that the deduction theorem holds for K_n'. Find similar (equivalent) variants of T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, and $\mathrm{KD45}_n$ for which the deduction theorem holds.

This shows that it is essentially R2—Knowledge Generalization—that causes the deduction theorem to fail for the logics that we have been considering.

3.9 Show that A6 is provable from A3, A1, and R1.

3.10 In this exercise, we consider when an agent can know both φ and $\neg\varphi$, or both φ and the fact that he does not know φ.

(a) Show that $K_1\varphi \wedge K_1\neg\varphi$ is consistent with K_n by constructing a Kripke structure that satisfies, for example, $K_1 p \wedge K_1 \neg p$.

(b) Show that $\mathrm{K}_n + \{A6\} \vdash \neg(K_i\varphi \wedge K_i\neg\varphi)$. (You may assume that $\mathrm{K}_n \vdash K_i(\varphi \wedge \psi) \Leftrightarrow (K_i\varphi \wedge K_i\psi)$; you are asked to prove this in Exercise 3.31(a).) Show as a consequence that $AX \vdash \neg(K_i\varphi \wedge K_i\neg\varphi)$ if AX is any one of T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, or $\mathrm{KD45}_n$.

(c) Show that $AX \vdash \neg K_i(\varphi \wedge \neg K_i\varphi)$ although $\varphi \wedge \neg K_i\varphi$ is consistent with AX, where AX is any of T_n, $\mathrm{S4}_n$, $\mathrm{S5}_n$, or $\mathrm{KD45}_n$. Thus, although it is consistent in each of these logics for φ to be true but agent i not to know it, it is impossible for i to know this fact.

*** 3.11** Give syntactic proofs of the following properties of common knowledge:

(a) $\mathrm{K}_n^C \vdash (C_G\varphi \wedge C_G(\varphi \Rightarrow \psi)) \Rightarrow C_G\psi$,

(b) $\mathrm{T}_n^C \vdash C_G\varphi \Rightarrow \varphi$,

(c) $\mathrm{K}_n^C \vdash C_G\varphi \Rightarrow C_G C_G\varphi$ (note that the analogous axiom A4 is *not* needed),

(d) $\mathrm{S5}_n^C \vdash \neg C_G\varphi \Rightarrow C_G\neg C_G\varphi$ (hint: show $\mathrm{S5}_n^C \vdash \neg C_G\varphi \Leftrightarrow K_i\neg C_G\varphi$ for all $i \in G$),

(e) $\mathrm{S4}_n^C \nvdash \neg C_G\varphi \Rightarrow C_G\neg C_G\varphi$ (hint: show that $\neg C_G\varphi \wedge \neg C_G\neg C_G\varphi$ is satisfiable in some structure in $\mathcal{M}_n^{rt}$),

(f) $\mathrm{K}_n^C \vdash C_G\varphi \Rightarrow C_{G'}\varphi$ if $G \supseteq G'$.

3.12 Prove Lemma 3.1.4.

3.13 In this exercise, we focus on the connection between axiom systems and possibility relations.

(a) Show that axiom A4 is valid in all structures in which the possibility relation is transitive.

(b) Show that axiom A5 is valid in all structures in which the possibility relation is Euclidean.

(c) Show that axiom A5 forces the possibility relation in the canonical structure to be Euclidean; specifically, show that if all instances of A5 are true at a state s_V in the canonical structure and (s_V, s_W), $(s_V, s_X) \in \mathcal{K}_i$, then $(s_W, s_X) \in \mathcal{K}_i$.

(d) Show that axiom A6 is valid in all structures in which the possibility relation is serial.

(e) Show that axiom A6 forces the possibility relation in the canonical structure to be serial; in particular, show that if all instances of A6 are true at a state s_V, then there must be some state s_W such that $(s_V, s_W) \in \mathcal{K}_i$.

*** 3.14** In this exercise, we show that the formulas proved valid in Exercise 2.12 are provable in $\mathrm{S5}_n$. Give syntactic proofs of the following properties:

(a) $\mathrm{S5}_n \vdash \neg\varphi \Rightarrow K_i\neg K_i\varphi$,

(b) $S5_n \vdash \neg\varphi \Rightarrow K_{i_1} \ldots K_{i_k} \neg K_{i_k} \ldots K_{i_1} \varphi$ for any sequence $i_1, \ldots, i_k$ of agents,

(c) $S5_n \vdash \neg K_i \neg K_i \varphi \Leftrightarrow K_i \varphi$.

(Hint: for part (b), use part (a), induction, the Knowledge Generalization Rule, and the Distribution Axiom.)

3.15 Prove that the structure M in Figure 3.1 is a model of $S5_1$. (Hint: show that there is a Kripke structure M' with a single state s' such that for every formula $\varphi \in \mathcal{L}_1(\{p\})$ we have $(M, s) \models \varphi$ iff $(M, t) \models \varphi$ iff $(M', s') \models \varphi$.)

3.16 In this exercise, we show that there is a construction that converts a model M of T_n (resp., $S4_n$, $S5_n$, $KD45_n$) to a model M' in $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$, $\mathcal{M}_n^{elt}$) that in a precise sense is equivalent to M. Given a Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, let $M^r = (S, \pi, \mathcal{K}_1^r, \ldots, \mathcal{K}_n^r)$, where $\mathcal{K}_i^r$ is the reflexive closure of $\mathcal{K}_i$; that is, $\mathcal{K}_i^r = \mathcal{K}_i \cup \{(s, s) \mid s \in S\}$. Similarly, let M^{rt} (resp., M^{rst}, M^{et}) be the structure obtained from M by replacing the $\mathcal{K}_i$ relations by their reflexive, transitive closures (resp., reflexive, symmetric and transitive closures; Euclidean and transitive closures). Note that we have M^{et} rather than M^{elt}, since it does not make sense to take the serial closure.

Prove the following:

(a) $M^r \in \mathcal{M}_n^r$; and if M is a model of T_n, then $(M, s) \equiv (M^r, s)$ for all states s in M.

(b) $M^{rt} \in \mathcal{M}_n^{rt}$; and if M is a model of $S4_n$, then $(M, s) \equiv (M^{rt}, s)$ for all states s in M.

(c) $M^{rst} \in \mathcal{M}_n^{rst}$; and if M is a model of $S5_n$, then $(M, s) \equiv (M^{rst}, s)$ for all states s in M.

(d) If M is a model of $KD45_n$, then so is M^{et}; moreover, in this case, $M^{et} \in \mathcal{M}_n^{elt}$ and $(M, s) \equiv (M^{et}, s)$ for all states s in M. Note that this case is slightly different from the previous cases, since it is not necessarily true in general that $M^{et} \in \mathcal{M}_n^{elt}$.

3.17 Let $\mathcal{F}_n$ be the class of all Kripke frames. Just as for structures, we can consider subclasses of $\mathcal{F}_n$ such as $\mathcal{F}_n^r$, $\mathcal{F}_n^{rt}$, $\mathcal{F}_n^{rst}$, and $\mathcal{F}_n^{elt}$. We say that a frame F is a *model of* T_n (resp., $S4_n$, $S5_n$, $KD45_n$) if every structure based on F is a model of T_n (resp., $S4_n$, $S5_n$, $KD45_n$). Prove the following:

(a) F is a model of T_n iff $F \in \mathcal{F}_n^r$,

(b) F is a model of $\mathrm{S4}_n$ iff $F \in \mathcal{F}_n^{rt}$,

(c) F is a model of $\mathrm{S5}_n$ iff $F \in \mathcal{F}_n^{rst}$,

(d) F is a model of $\mathrm{KD45}_n$ iff $F \in \mathcal{F}_n^{elt}$.

3.18 In this exercise, we take a closer look at axiom A7.

(a) Show that axiom A7 is provable from the system consisting of A1, A3, A5, and R1.

(b) Show that axiom A7 forces the possibility relation in the canonical structure to be symmetric.

*** 3.19** Show that A4 is provable from the system consisting of A1, A2, A3, A5, R1, and R2. (Hint: use Exercise 3.18 to show that $K_i\varphi \Rightarrow K_i\neg K_i\neg K_i\varphi$ is provable, and use A5 and propositional reasoning to show that $\neg K_i\neg K_i\varphi \Rightarrow K_i\varphi$ is also provable.)

3.20 Prove, using Lemma 3.1.4 and the techniques of Theorem 3.1.5, that the following axiom systems are equivalent (i.e., precisely the same formulas are provable in all of these systems):

(a) $\mathrm{S5}_n$,

(b) the system consisting of {A1,A2,A4,A6,A7,R1,R2},

(c) the system consisting of {A1,A2,A3,A5,R1,R2},

(d) the system consisting of {A1,A2,A3,A4,A7,R1,R2}.

(Note that this exercise gives us an indirect proof of the preceding exercise.)

3.21 Fill in the missing details in the proof of Proposition 3.1.6. In particular, show that the relation $\mathcal{K}_1'$ defined in the the proof of part (b) has the properties claimed for it, and show that $(M, s) \equiv (M', s)$ for all states $s \in \{s_0\} \cup \mathcal{K}_1'(s_0)$ and all formulas ψ.

3.22 Show that an analogue to Proposition 3.1.6(b) holds for K45. (Hint: the only difference is that we can now take the set S to be empty.)

*** 3.23** The *depth* of a formula is the depth of nesting of the K_i operators in the formula. Formally, we define depth by induction on structure of formulas. We define $depth(p) = 0$ for a primitive proposition p, $depth(\neg\varphi) = depth(\varphi)$, $depth(\varphi \wedge \psi) = \max(depth(\varphi), depth(\psi))$, and $depth(K_i\varphi) = depth(\varphi) + 1$.

(a) Show that for every formula $\varphi \in \mathcal{L}_1$ we can effectively find a formula φ' of depth 1 such that S5 $\vdash \varphi \Leftrightarrow \varphi'$. That is, for every formula in $\mathcal{L}_1$ we can effectively find an equivalent formula that is a Boolean combination of propositional formulas and formulas of the form $K_1\psi$, where ψ is propositional. (Hint: use the fact that $K_1(\varphi_1 \vee K_1\varphi_2) \Leftrightarrow (K_1\varphi_1 \vee K_1\varphi_2)$ is valid in $\mathcal{M}_1^{rst}$.)

(b) Show that for every formula in $\mathcal{L}_1$ of the form $K_1\varphi$ we can effectively find an equivalent formula that is a Boolean combination of formulas of the form $K_1\psi$, where ψ is propositional.

3.24 Extend Proposition 3.2.1 to deal with formulas in the language $\mathcal{L}_n^{CD}$. (We remark that once we have common knowledge in the language, the algorithm will no longer run in time $O(||M|| \times |\varphi|)$, but will still run in time polynomial in $||M||$ and $|\varphi|$.)

**** 3.25** In this exercise, we sketch the details of how to construct effectively the proof of a valid formula. (By "construct effectively," we mean that there is an algorithm that takes as input a formula φ and gives as output a proof of φ, if φ is valid, and halts, say with the output "not valid," if φ is not valid.) We work with K_n here, but the proof can be easily modified to deal with all the other logics we have been considering. Using the notation of Theorem 3.2.2, let $Sub^+(\varphi)$ consist of all the subformulas of φ and their negations. Let $\mathcal{V}$ consist of all subsets V of $Sub^+(\varphi)$ such that (a) $\psi \in V$ iff $\neg\psi \notin V$ for each subformula ψ of φ and (b) $\psi \wedge \psi' \in V$ iff $\psi, \psi' \in V$. Let $M^0 = (S^0, \pi^0, \mathcal{K}_1^0, \ldots, \mathcal{K}_n^0)$, where $S^0 = \{s_V \mid V \in \mathcal{V}\}$ and $\pi^0, \mathcal{K}_1^0, \ldots, \mathcal{K}_n^0$ are constructed as in the proof of Theorem 3.1.3. We construct inductively, for $k = 0, 1, 2, \ldots$, a sequence of structures $M^k = (S^k, \pi^k, \mathcal{K}_1^k, \ldots, \mathcal{K}_n^k)$. Suppose that we have already constructed M^k. Let S^{k+1} consist of those states $s_V \in S^k$ such that if there is a formula of the form $\neg K_i\psi \in V$, then there is a state $s_W \in S^k$ such that $(s_V, s_W) \in \mathcal{K}_i^k$ and $\neg\psi \in W$. Let $\pi^{k+1}, \mathcal{K}_1^{k+1}, \ldots, \mathcal{K}_n^{k+1}$ be the restrictions of $\pi^k, \mathcal{K}_1^k, \ldots, \mathcal{K}_1^k$ to S^{k+1}. Note that this construction is effective. Moreover, since $|S^0| \leq 2^{|\varphi|}$ and $S^{k+1} \subseteq S^k$, there must be some point, say k_0, such that $S^{k_0+1} = S^{k_0}$.

(a) Prove that φ is satisfiable iff for some state $s_V \in S^{k_0}$ we have $\varphi \in V$. (Note that this gives us another proof that if φ is satisfiable, it is satisfiable in a finite structure.)

(b) Let φ_V be the conjunction of all the formulas in the set V. Prove that if $s_V \in (S^k - S^{k+1})$, then $\mathrm{K}_n \vdash \neg\varphi_V$; moreover, show that the proof of $\neg\varphi_V$ can be effectively constructed. (Hint: show by induction on k that $\mathrm{K}_n \vdash \bigvee_{s_V \in S^k} \varphi_V$, and that the proof of $\bigvee_{s_V \in S^k} \varphi_V$ can be constructed effectively.)

(c) Note that if φ is valid, then if we apply the previous construction to $\neg\varphi$, eventually we eliminate every state s_V such that $\neg\varphi \in V$. Use this observation and parts (a) and (b) to show that we can effectively construct a proof of φ.

3.26 Complete the details of the proof of Theorem 3.2.4.

3.27 In the proof of Theorem 3.3.1, show that $|S_\varphi| \leq 2^{|\varphi|}$. (Hint: recall that $|C_G\varphi| = 2 + 2\,|G| + |\varphi|$.)

*** 3.28** In this exercise, we fill in some of the details of the proof of Claim 3.1 in the proof of Theorem 3.3.1. Assume that $(M_\varphi, s_V) \models C_G\psi$. Let $\mathcal{W}$ be as in the proof of Theorem 3.3.1. We wish to prove that

$$\mathrm{K}_n^C \vdash \varphi_{\mathcal{W}} \Rightarrow E_G(\psi \wedge \varphi_{\mathcal{W}}).$$

(a) Prove that if $i \in G$ and $W \in \mathcal{W}$, then $\mathrm{K}_n^C \vdash \varphi_W \Rightarrow K_i\psi$. (Hint: assume that $W/K_i = \{\varphi_1, \ldots, \varphi_k\}$. Use an argument like that in the proof of Theorem 3.1.3, where W here plays the role that V plays there, and use the fact that $(M_\varphi, s_W) \models K_i\psi$, to show that

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

Now use the fact that $K_i\varphi_j \in W$, for $j = 1, \ldots, k$.)

(b) Define $\overline{\mathcal{W}}$ to be $Con_C(\varphi) - \mathcal{W}$. Show that if $i \in G$, $W \in \mathcal{W}$, and $W' \in \overline{\mathcal{W}}$, then $\mathrm{K}_n^C \vdash \varphi_W \Rightarrow K_i\neg\varphi_{W'}$. (Hint: by definition of $\mathcal{W}$, show that $(M_\varphi, s_W) \models C_G\psi$ and $(M_\varphi, s_{W'}) \not\models C_G\psi$. Conclude that $s_{W'}$ is not G-reachable from s_W and, in particular, $(s_W, s_{W'}) \notin \mathcal{K}_i$. By definition of $\mathcal{K}_i$, conclude that $W/K_i \not\subseteq W'$, so there is a formula ψ' such that $K_i\psi' \in W$ and $\psi' \notin W'$. Since $\psi' \notin W'$, show that $\mathrm{K}_n^C \vdash \psi' \Rightarrow \neg\varphi_{W'}$. From this, show $\mathrm{K}_n^C \vdash K_i\psi' \Rightarrow K_i\neg\varphi_{W'}$. Since $K_i\psi' \in W$, conclude that $\mathrm{K}_n^C \vdash \varphi_W \Rightarrow K_i\neg\varphi_{W'}$.)

(c) Conclude from parts (a) and (b) that

$$\mathrm{K}_n^C \vdash \varphi_W \Rightarrow K_i\Big(\psi \wedge \Big(\bigwedge_{W' \in \overline{\mathcal{W}}} \neg\varphi_{W'}\Big)\Big).$$

(d) Prove that

$$\mathrm{K}_n^C \vdash \varphi_{\mathcal{W}} \Leftrightarrow \Big(\bigwedge_{W' \in \overline{\mathcal{W}}} \neg\varphi_{W'}\Big).$$

(Hint: first show that $\mathrm{K}_n^C \vdash \bigvee_{W \in Con_C(\varphi)} \varphi_W$.)

(e) Use parts (c) and (d) to show that $\mathrm{K}_n^C \vdash \varphi_W \Rightarrow K_i(\psi \wedge \varphi_{\mathcal{W}})$.

(f) Conclude from part (e) that $\mathrm{K}_n^C \vdash \varphi_{\mathcal{W}} \Rightarrow E_G(\psi \wedge \varphi_{\mathcal{W}})$.

*** 3.29** This exercise provides a weak form of the deduction theorem for languages with common knowledge. Let $G = \{1, \ldots, n\}$. Show that if $\mathrm{K}_n^C, \varphi \vdash \psi$, then $\mathrm{K}_n^C \vdash C_G\varphi \Rightarrow C_G\psi$. Observe that similar results hold for T_n^C, $\mathrm{S4}_n^C$, $\mathrm{S5}_n^C$, and $\mathrm{KD45}_n^C$.

*** 3.30** In this exercise, we fill in some details of the proof of Theorem 3.4.1.

(a) Show that $\mathcal{K}_G \subseteq \bigcap_{i \in G} \mathcal{K}_i$.

(b) Construct an example where $\mathcal{K}_G \neq \bigcap_{i \in G} \mathcal{K}_i$.

(c) Show that the canonical structure (or any other structure for that matter) can be *unwound* to get a structure whose graph looks like a tree, in such a way that the same formulas are true in corresponding states. (More formally, given a structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, there is another structure $M' = (S', \pi', \mathcal{K}'_1, \ldots, \mathcal{K}'_n)$ and a function $f : S' \to S$ such that (i) the graph of M' looks like a tree, in that for all states s', t' in M', there is at most one path from s' to t', and no path from s' back to itself, (ii) if $(s', t') \in \mathcal{K}'_i$ then $(f(s'), f(t')) \in \mathcal{K}_i$, (iii) $\pi'(s') = \pi(f(s'))$, and (iv) f is onto, so that for all $s \in S$ there exists $s' \in S'$ such that $f(s') = s$. Moreover, we have $(M', s') \models \varphi$ iff $(M, f(s')) \models \varphi$ for all states $s' \in S'$ and all formulas $\varphi \in \mathcal{L}_n^D$.)

(d) Show that we can unwind the canonical structure in such a way as to get a structure M' where $\mathcal{K}_G = \bigcap_{i \in G} \mathcal{K}_i$.

3.31 Prove from the axioms that knowledge distributes over conjunctions. That is, give syntactic proofs of the following:

(a) $\mathrm{K}_n \vdash K_i(\varphi \wedge \psi) \Leftrightarrow (K_i\varphi \wedge K_i\psi)$ (hint: use the observation that $\varphi \Rightarrow (\psi \Rightarrow (\varphi \wedge \psi))$ is a propositional tautology),

(b) $\mathrm{K}_n^C \vdash E_G(\varphi \wedge \psi) \Leftrightarrow (E_G\varphi \wedge E_G\psi)$,

(c) $\mathrm{K}_n^C \vdash C_G(\varphi \wedge \psi) \Leftrightarrow (C_G\varphi \wedge C_G\psi)$,

(d) $\mathrm{K}_n^D \vdash D_G(\varphi \wedge \psi) \Leftrightarrow (D_G\varphi \wedge D_G\psi)$.

3.32 In Chapter 2, we said that distributed knowledge could be viewed as the knowledge the agents would have by pooling their individual knowledge together. This suggests the following inference rule:

RD1. From $(\psi_1 \wedge \ldots \wedge \psi_k) \Rightarrow \varphi$ infer $(K_{i_1}\psi_1 \wedge \ldots \wedge K_{i_k}\psi_k) \Rightarrow D_G\varphi$, for $G = \{i_1, \ldots, i_k\}$.

Intuitively, RD1 says that if $\psi = \psi_1 \wedge \ldots \wedge \psi_k$ implies φ, and if each of the agents in G knows a "part" of ψ (in particular, agent i_j knows ψ_j), then together they have distributed knowledge of ψ, and thus distributed knowledge of φ.

(a) Prove that RD1 preserves validity with respect to $\mathcal{M}_n$.

(b) Show that RD1 is derivable from axiom A2 (with D_G substituted for K_i), D1, and D2, using propositional reasoning. (Hint: you will also need the results of Exercise 3.31.)

3.33 We say that φ is a *pure knowledge formula* if φ is a Boolean combination of formulas of the form $K_i\psi$ (that is, it is formed from formulas of the form $K_i\psi$ using $\wedge$, $\neg$, and $\vee$). For example, $K_2p \vee (K_1\neg K_3p \wedge \neg K_2\neg p)$ is a pure knowledge formula, but $p \wedge \neg K_1p$ is not. Show that if φ is a pure knowledge formula, then $\mathrm{K}_n^D \vdash \varphi \Rightarrow D_G\varphi$.

3.34 Fill in the details of the proofs of Proposition 3.6.3 and Theorem 3.6.4.

3.35 Prove analogues to Proposition 3.6.3 and Theorem 3.6.4 for K45. (Hint: use Exercise 3.22.)

*** 3.36** Show that $K_1\varphi$ is S4-consistent iff $K_1\varphi$ is S5-consistent. Conclude that the satisfiability problem for S4 for formulas of the form $K_1\varphi$ is *NP*-complete (although the general satisfiability problem for S4 is *PSPACE*-complete).

*** 3.37** In this exercise, we show that first-order logic is, in a precise sense, expressive enough to capture propositional modal logic (without common knowledge). Given a set Φ of primitive propositions, let the vocabulary Φ^* consist of a unary predicate P corresponding to each primitive proposition p in Φ, as well as binary predicates $R_1, \ldots, R_n$, one for each agent. We now define a translation from formulas in $\mathcal{L}_n(\Phi)$ to first-order formulas over Φ^*, so that for every formula $\varphi \in \mathcal{L}_n(\Phi)$ there is a corresponding first-order formula φ^* with one free variable x:

- $p^* = P(x)$ for a primitive proposition p
- $(\neg\varphi)^* = \neg(\varphi^*)$
- $(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^*$
- $(K_i\varphi)^* = \forall y(R_i(x, y) \Rightarrow \varphi^*(y))$, where y is a new variable not appearing in φ^* and $\varphi^*(y)$ is the result of replacing all occurrences of x in φ^* by y.

Next, we provide a mapping from a Kripke structure $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n) \in \mathcal{M}_n(\Phi)$ to a relational Φ^*-structure M^*. The domain of M^* is S. For each primitive proposition $p \in \Phi$, we let $P^{M^*} = \{s \in S \mid \pi(s)(p) = \textbf{true}\}$, and let $R_i^{M^*} = \mathcal{K}_i$.

(a) Show that $(M, s) \models \varphi$ iff $(M^*, V) \models \varphi^*(x)$, where $V(x) = s$. Intuitively, this says that φ^* is true of exactly the domain elements corresponding to states s for which $(M, s) \models \varphi$.

(b) Show that φ is valid with respect to $\mathcal{M}_n(\Phi)$ iff $\forall x\varphi^*(x)$ is a valid first-order formula. (Hint: use the fact that the mapping from structures in $\mathcal{M}_n(\Phi)$ to relational Φ^*-structures is invertible.)

(c) Show how to modify this construction to capture validity with respect to structures in $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$, $\mathcal{M}_n^{rst}$, $\mathcal{M}_n^{elt}$).

Given this translation, we might wonder why we should consider propositional modal logic at all. There are four main reasons for this. First, the syntax of modal logic allows us to more directly capture the types of statements regarding knowledge

that we typically want to make. Second, the semantics of modal logic in terms of possible-worlds structures better represents our intuitions (and, as we shall see, directly corresponds to a standard representation of a system, one of our major application areas). Third, the translation fails for common knowledge. (That is, there is no first-order formula corresponding to common knowledge. This follows from the fact that transitive closure cannot be expressed in first-order logic.) Finally, by moving to first-order logic, we lose the nice complexity properties that we have for propositional modal logic. First-order logic is undecidable; there is no algorithm that can effectively decide whether a first-order formula is valid.

3.38 Show that $(\mathcal{A}, V) \models \forall x\varphi$ iff $(\mathcal{A}, V[x/a]) \models \varphi$ for every $a \in dom(\mathcal{A})$.

3.39 Inductively define what it means for an occurrence of a variable x to be free in a formula as follows:

- if φ is an atomic formula ($P(t_1, \ldots, t_k)$ or $t_1 = t_2$), then every occurrence of x in φ is free,
- an occurrence of x is free in $\neg\varphi$ iff the corresponding occurrence of x is free in φ,
- an occurrence of x is free in $\varphi_1 \wedge \varphi_2$ iff the corresponding occurrence of x in φ_1 or φ_2 is free,
- an occurrence of x is free in $\exists y\varphi$ iff the corresponding occurrence of x is free in φ and x is different from y.

A *sentence* is a formula in which no occurrences of variables are free.

(a) Show that if φ is a formula, and V and V' are valuations that agree on all of the variables that are free in φ, then $(\mathcal{A}, V) \models \varphi$ iff $(\mathcal{A}, V') \models \varphi$.

(b) Show that if φ is a sentence, and V and V' are valuations on the structure $\mathcal{A}$, then $(\mathcal{A}, V) \models \varphi$ iff $(\mathcal{A}, V') \models \varphi$.

3.40 In this exercise, we consider a semantics without any assumptions whatsoever about relationships between domains of worlds within a relational Kripke structure. For simplicity, we assume that there are no function symbols. Given a relational Kripke structure M, we now take a valuation V on M to be a mapping from variables

to the union of the domains at the states of M. Then, as we saw, to define what it means for a formula such as $K_i Tall(x)$ to hold at a state s of a relational Kripke structure M under a valuation V such that $V(x) = Bill$, we may have to decide if $Tall(x)$ is true at a state t such that *Bill* is not in the domain of the relational structure $\pi(t)$. One solution to this problem is to note that if *Bill* is not in the domain of $\pi(t)$, then certainly $Bill \notin Tall^{\pi(t)}$. Therefore, we define a new semantics where we simply say $(M, t, V) \not\models Tall(x)$ if $V(x)$ is not in the domain of $\pi(t)$. Similarly, we say $(M, t, V) \not\models x = y$ if $V(x)$ or $V(y)$ is not in the domain of $\pi(t)$. Further, we modify our standard semantics by saying $(M, s, V) \models \exists x\varphi$ iff $(M, s, V[x/a]) \models \varphi$ for some $a \in dom(\pi(s))$. Although this semantics has some attractive features, it has its problems, as we now show.

The *universal closure* of a formula φ is the formula $\forall x_1 \ldots \forall x_k \varphi$, where $x_1, \ldots, x_k$ are all of the variables that occur free in φ. In first-order logic, it is easy to see that a formula is valid if and only if its universal closure is valid. The next two parts of the exercise, however, show that this is not the case for the semantics of this exercise.

(a) Show that $\forall x(x = x)$ is valid under the semantics of this exercise.

(b) Show that $x = x$ is not valid under the semantics of this exercise.

The fact that $x = x$ is not valid in this semantics is certainly undesirable. Of course, the formula $x = x$ is not a sentence. The next part of this exercise gives an example of a sentence that is not valid in this logic, which we might hope would be valid, namely, the universal closure of the Knowledge of Equality axiom.

(c) Show that the formula $\forall x K_i(x = x)$ is not valid.

The failure of formulas such as those in (b) and (c) to be valid have led most researchers to reject this semantics as a general solution.

We might hope to solve this problem by redefining $(M, t, V) \models (x = y)$ iff $V(x) = V(y)$, irrespective of whether x or y is in domain of $\pi(t)$. While this change "solves" the problems of (b) and (c), other problems remain.

(d) Show that under this redefinition, neither $\exists y(x = y)$ nor $\forall x K_i \exists y(x = y)$ is valid.

The problems that arise in part (d) are due to the fact that $\exists x\varphi$ is true at s if φ holds for some a in $dom(\pi(s))$. We could solve this problem by taking $\exists x\varphi$ to hold if φ

holds for any a in the union of the domains at all states. This indeed solves all the problems we have raised but effectively puts us back in the common-domain setting. The semantics is now equivalent to that which would be obtained by taking the same domain at all states, namely, the union of all the domains.

3.41 Show that K_n is sound for relational Kripke structures with n agents. Show that if each $\mathcal{K}_i$ is an equivalence relation, then $\mathrm{S5}_n$ is sound.

3.42 In this exercise, we consider when axioms (3.3) and (3.4) from Section 3.7.4 are valid.

(a) Show that both axioms are valid with respect to relational structures.

(b) We say that a constant, function, or relation symbol is a *rigid designator* if it takes on the same value in every state. We say that a term is a rigid designator if all the constant and function symbols that appear in it are rigid designators. Show that both axioms are valid with respect to relational Kripke structures if t, t_1, and t_2 are rigid designators.

We remark that in certain applications it may be useful to designate some of the symbols as rigid designators, while others are allowed to vary. For example, we may want the interpretation of constants such as 0 and 1 and of functions such as $+$ and $\times$ to be independent of the state.

*** 3.43** In this exercise, we consider the Barcan formula.

(a) Show that the Barcan formula is valid.

(b) Show that this axiom is a consequence of the axioms and rules of $\mathrm{S5}_n$ together with the axioms and rules of first-order logic. (Hint: first-order logic has analogues to the Distribution Axiom A2 and the Knowledge Generalization Rule R2 for universal quantification: $(\forall x\varphi \wedge \forall x(\varphi \Rightarrow \psi)) \Rightarrow \forall x\psi$ is an axiom, and "from φ infer $\forall x\varphi$" is an inference rule. In addition, there is the following axiom:

$$\varphi \Rightarrow \forall x\varphi \text{ if } \varphi \text{ has no free occurrences of } x.$$

Using these, the restricted version of (3.3), and the axioms of $\mathrm{S5}_n$, prove

$$\neg K_i \neg \forall x_1 \ldots \forall x_k K_i \varphi \Rightarrow \forall x_1 \ldots \forall x_k \varphi.$$

Then show that, using the axioms and rules of S5$_n$, from $\neg K_i \neg \psi_1 \Rightarrow \psi_2$ we can prove $\psi_1 \Rightarrow K_i \psi_2$.)

3.44 Show that under the domain-inclusion assumption

(a) the Barcan formula is not valid,

(b) the converse of the Barcan formula, namely

$$K_i \forall x_1 \ldots \forall x_k \varphi \Rightarrow \forall x_1 \ldots \forall x_k K_i \varphi,$$

is valid,

(c) if the $\mathcal{K}_i$ relations are equivalence relations, then the Barcan formula is valid.

3.45 In this exercise, we consider the Knowledge of Inequality axiom.

(a) Show that the Knowledge of Inequality axiom (3.8) is valid.

(b) Show that this axiom is a consequence of the axioms and rules of S5$_n$ together with the axioms and rules of first-order logic. (Hint: show that $K_i(\varphi \Rightarrow K_i \varphi) \Rightarrow K_i(\neg\varphi \Rightarrow K_i \neg\varphi)$ is valid in S5$_n$, and hence provable in S5$_n$. Now take φ to be $x_1 = x_2$, and apply Knowledge of Equality.)

Notes

A discussion of different varieties of modal logic can be found in some of the standard texts in the area, such as [Hughes and Cresswell 1996], [Chellas 1980], and [Blackburn, de Rijke, and Venema 2001]. The historical names S4 and S5 are due to Lewis, and are discussed in his book with Langford [1959]. The names K and T are due to Lemmon [1977], as is the idea of naming the logic for the significant axioms used. Arguments for using logics weaker than S5 in game theory can be found in, for example, [Samet 1987] and [Geanakoplos 1989].

The treatment of completeness and complexity issues in this chapter largely follows that of [Halpern and Moses 1992]. The technique for proving completeness using canonical structures seems to have been worked out independently by Makinson

[1966], Kaplan [1966], and Lemmon and/or Scott [Lemmon 1977]. An algebraic approach to the semantics of modal logic is described by Lemmon [1977]. Frames were introduced by Lemmon and Scott [Lemmon 1977], who called them "world systems." The term "frame" is due to Segerberg [1968]. The idea of using frames to characterize axiom systems (as in Exercise 3.17) is well known in modal logic; it appears, for example, in [Goldblatt 1992] and [Hughes and Cresswell 1984].

Although we restrict our attention in this book to languages $\mathcal{L}$ with a countable set of formulas, this is not really necessary. For example, we make this restriction in Lemma 3.1.2 only to simplify the proof. Indeed, Lemma 3.1.2 is a standard result in the model-theoretic literature and is known as *Lindenbaum's Theorem* [Chang and Keisler 1990, Proposition 1.3.11].

As we mentioned in the notes to Chapter 1, Lenzen's overview article [1978] has a good discussion and review of philosophers' arguments for and against various axioms of knowledge. In the next chapter we present our model of knowledge in multi-agent systems for which $S5_n$ is an appropriate axiomatization. Other axiom systems for knowledge have been used in various contexts. Moore [1985] uses S4 in his theory of knowledge and action. Since the knowledge represented in a knowledge base is typically not required to be true, axiom A3 has been thought inappropriate for these applications; thus KD45 is considered, for example, by Levesque [1984a]. KD45 has also been considered, for example, by Fagin and Halpern [1988a] and by Levesque [1984b], to be an appropriate logic for characterizing the beliefs of an agent, who might believe things that in fact turn out to be false.

There has been a great deal of interest recently in having a system with modal operators for knowledge and belief where, typically, the belief operator satisfies the axioms of KD45 and the knowledge operator satisfies the axioms of S5. The focus has been on the interaction between these operators (for example, if agent i believes φ, does she know that she believes φ?) and on defining belief in terms of knowledge. Further details can be found in [Friedman and Halpern 1997], [Kraus and Lehmann 1988], [Moses and Shoham 1993], and [Voorbraak 1992].

The formula $K_i(\varphi \wedge \neg K_i\varphi)$ discussed in part (c) of Exercise 3.10 has been called a "pragmatically paradoxical formula." It was first introduced by Moore (see [Hintikka 1962]).

Axioms for common knowledge appear in [Lehmann 1984], [Milgrom 1981], and [McCarthy, Sato, Hayashi, and Igarishi 1979]. (In these papers, only the modal operator C, referring to common knowledge of all the agents in the system, was used, rather than the indexed modal operator C_G.) The essential ideas for extending the

canonical structure technique to languages including common knowledge are due to Kozen and Parikh [1981], who proved completeness results for the logic *PDL* (Propositional Dynamic Logic) in this way. The idea for proving completeness for the language including distributed knowledge is due to Halpern and Moses [1992]; a formal completeness proof (as in Exercise 3.30) can be found in [Fagin, Halpern, and Vardi 1992a] and [Hoek and Meyer 1992].

An excellent introduction to complexity theory is given by Hopcroft and Ullman [1979]. The fact that satisfiability for propositional logic is *NP*-complete was proved by Cook [1971], who in fact introduced the notions of *NP* and *NP*-completeness. Ladner [1977] proved that the satisfiability problem for S5 is *NP*-complete, and that satisfiability for the logics K, T, and S4 is *PSPACE*-complete. The results in the multi-agent case are from [Halpern and Moses 1992]. The exponential time results for logics involving common knowledge are based on similar results for PDL. The lower bound for PDL is due to Fischer and Ladner [1979]; the matching upper bound is due to Pratt [1979]. Details of the proofs of the complexity results not included here can be found in [Halpern and Moses 1992]. A general framework for studying the complexity of modal logics is described by Vardi [1989]. For a recent overview of the complexity of modal logics, see [Blackburn, de Rijke, and Venema 2001].

An excellent introduction to first-order logic is [Enderton 1972]; this book also provides a nice discussion of issues of decidability and undecidability. The translation from modal logic to first-order logic (Exercise 3.37) is another notion that seems to have been developed independently by a number of people. The first treatment of these ideas in print seems to be due to van Benthem [1974]; details and further discussion can be found in his book [1985]. The distinction between "knowing that" and "knowing who" is related to an old and somewhat murky distinction between knowledge *de dicto* (literally, "knowledge of words") and knowledge *de re* (literally, "knowledge of things"). Plantinga [1974] discusses these terms in more detail.

The example of the morning star and the evening star is due to Frege [1892], and its implications for first-order modal logic were first discussed by Quine [1947]. The idea of dealing with the morning-star paradox by restricting substitutions so that they are not allowed within the scope of knowledge operators K_i is due to Kanger [1957a], and the idea of using rigid designators is due to Kaplan [1969].

The Barcan formula (or actually, a formula equivalent to it) was introduced by Barcan [1946]. Prior [1956] showed that the Barcan formula is a consequence of the axioms of first-order logic, along with S5; see [Hughes and Cresswell 1968, page 145] for a proof. Prior [1957] also made an early objection to it. Kripke

[1963b] introduced structures equivalent to relational Kripke structures, but where the domains of distinct worlds can be unrelated, and so the Barcan formula is violated. He gave a completeness proof for a first-order modal logic in the S5 case [1959]. Barcan [1947] showed the validity of the Knowledge of Equality axiom.

Detailed discussions of first-order modal logic, along with completeness proofs, appear in Hughes and Cresswell's book [1968] and Garson's article [1984]. Much of the information we have given about first-order modal logic (including bibliographic references) is from Hughes and Cresswell's book. They, along with Garson, discuss and prove sound and complete axiomatizations under a variety of assumptions, including cases where formulas involving equality are not allowed. Garson discusses in detail a number of ways of dealing with what is called the problem of "quantifying-in": how to give semantics to a formula such as $\exists x K_i(P(x))$ without the common domain assumption.

An axiomatization of first-order logic that includes (3.3) and a slightly stronger version of (3.4) appear in [Enderton 1972]. This stronger version says that if φ' is the result of replacing some occurrences of t_1 in $\varphi(t_1)$ by t_2, then $(t_1 = t_2) \Rightarrow (\varphi(t_1) \Leftrightarrow \varphi(t_2))$. It is not hard to show that in the presence of the other axioms, this stronger version is implied by our (3.4). We remark that in [Enderton 1972], the Rule of Universal Generalization ("from φ infer $\forall x\varphi$") is not used. Instead, all axioms are viewed as universally quantified. We do assume this rule here. (Alternatively, we would have to universally quantify all the free variables in the axioms of K_n or $\mathrm{S5}_n$.)