2

Epistemic Logic

2.1 Introduction

In this chapter we introduce the basic epistemic logic, to which we add a dynamic component in consecutive chapters. Epistemic logic, as it is conceived today, is very much influenced by the development of modal logic, and, in particular, by its Kripke semantics. We will emphasise the intuitive appeal of this semantics in this chapter and, indeed, throughout the book, since also the dynamics of epistemics will fruitfully utilise them.

The logical system *S5* is by far the most popular and accepted epistemic logic, and we will without further ado present it in this chapter, as a basis for the rest of the book. We do this first for agents in a group each with their own individual knowledge (Section 2.2), and then look at group notions of knowledge (Section 2.3), most notably that of common knowledge. On the fly, we give several examples and exercises.

Most, if not all of the material covered here belongs to the 'core of folklore in epistemic logic', therefore, the emphasis is on semantics and concepts, rather than on proofs of theorems.

Having presented the basic material, we then in Section 2.4 comment upon how the material in this chapter carries over to the notion of belief, rather than knowledge, and, finally, a section called 'Notes' (Section 2.5), collects the bibliographical and other meta-information regarding this chapter.

2.2 Basic System: *S5*

We now present the basic system for knowledge for a group of agents. Doing so, we follow a traditional outline, presenting the language, semantics, and axiomatisation in subsequent subsections.

2.2.1 Language

The basic language for knowledge is based on a countable set of atomic propositions P and a finite set of (names for) agents, A. Atomic propositions $p, q, \ldots$ describe some state of affairs in 'the actual world', or in a game, for example. In the following, p is an arbitrary atomic proposition from P, and a denotes an arbitrary agent from A.

Definition 2.1 (Basic epistemic language) Let P be a set of atomic propositions, and A a set of agent-symbols. The language $\mathcal{L}_K$, the language for multi-agent epistemic logic, is generated by the following BNF:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \qquad \square$$

This BNF-notation says that atoms are formulas, and one can build complex formulas from formulas using negation '$\neg$', conjunction '$\wedge$' and knowledge operators 'K_a'. Examples of formulas that can be generated using this definition are $(p \wedge \neg K_a K_a(p \wedge K_a \neg p))$ and $\neg K_a \neg(p \wedge K_a(q \wedge \neg K_a r))$. Here, p is an atom from P and a is an agent name from A. Other typical elements that we will use for atoms are q, r, and also p', q', r', $\ldots p''$, $\ldots$. Other variables for agents will be b, c and a', b', $\ldots$. In examples, we will also use other symbols for the agents, like S for Sender, R for receiver, etc. Throughout the book, we will use a number of 'standard abbreviations and conventions', like $(\varphi \vee \psi) = \neg(\neg\varphi \wedge \neg\psi)$, the symbol $\top$ as an abbreviation for $p \vee \neg p$ (for an arbitrary $p \in P$) and $\bot$ denoting $\neg\top$. Moreover, $(\varphi \rightarrow \psi) = (\neg\varphi \vee \psi)$, and $(\varphi \leftrightarrow \psi)$ is a shorthand for the conjunction of the implication in both directions. We omit outermost parenthesis if doing so does not lead to confusion.

For every agent a, $K_a\varphi$ is interpreted as "agent a knows (that) φ". Note that the epistemic language is a rather simple extension of that of propositional logic: we just add a unary operator for every agent. Here, an agent may be a human being, a player in a game, a robot, a machine, or simply a 'process'. We also introduce some epistemic definitions. The fact that a does not know that $\neg\varphi$ ($\neg K_a \neg\varphi$) is sometimes also pronounced as: "φ is consistent with a's knowledge", or, "the agent considers it possible that φ" (cf. also Section 2.2.2). We write $\hat{K}_a\varphi$ for this: $\hat{K}_a\varphi = \neg K_a \neg\varphi$. For any group B of agents from A, "Everybody in B knows φ", written $E_B\varphi$, is defined as the conjunction of all individuals in B knowing φ. Thus, we add an E_B operator for every $B \subseteq A$:

$$E_B\varphi = \bigwedge_{b \in B} K_b\varphi$$

Analogously to $\hat{K}$, we define $\hat{E}_B\varphi$ as $\neg E_B \neg\varphi$. Using the definition of E_B, this unravels to $\bigvee_{b \in B} \hat{K}_b\varphi$: "at least one individual in the group B considers φ a possibility".

Hence, examples of well-formed formulas are $p \wedge \neg K_a p$ ("p is true but agent a does not know it") and $\neg K_b K_c p \wedge \neg K_b \neg K_c p$ (saying that "agent b does not know *whether* agent c knows p). Another example is $K_a(p \rightarrow E_B p)$ ("agent a knows, that if p is true, everybody in B will know it").

Exercise 2.2 We have three agents, say a (Anne), b (Bill), and c (Cath). We furthermore have two atoms, p ("Anne has a sister") and q ("Anne has a brother"). Translate the following expressions in our formal language:

1. If Anne had a sister, she would know it.
2. Bill knows that Anne knows whether she has a sister.
3. Cath knows that if Anne has a sibling, it is a sister.
4. Anne considers it possible that Bill does not know that Anne has a sister.
5. Everybody in the group of three knows that Anne does not have a sibling if she does not know to have one.
6. Anne knows that if there is anybody who does not know that she has a sister, it must be Bill. □

We will often refer to $K_a, K_b, \ldots, \hat{K}_a, \hat{K}_b, \ldots, E_B, \hat{E}_B$ as *epistemic operators*, or sometimes, and more generally, as *modal operators*. For any modal operator X, we define $X^0\varphi$ to be equal to φ, and $X^{n+1}\varphi$ to be $XX^n\varphi$. We will also apply this convention to sequences of formulas. Hence, for instance $E_A^2\varphi$ means that everybody knows that everybody knows φ, and $(K_a\hat{K}_b)^2\varphi$ says that a knows that b considers it possible that a knows that b considers it possible that φ.

To demonstrate the usefulness of a language in which we allow iterations of individual epistemic operators, we now look at a simple *protocol specification* using epistemic operators. The derivation and correctness proofs of such protocols were a main motivation for computer scientists to study epistemic logic.

Example 2.3 (Alternating bit protocol) There are two processors, let us say a 'Sender S' and a 'Receiver R', although this terminology is a little bit misleading, since both parties can send messages to each other. The goal is for S to read a tape $X = \langle x_0, x_1, \ldots \rangle$, and to send all the inputs it has read to R over a communication channel. R in turn writes down everything it receives on an output tape Y. Unfortunately the channel is not trustworthy, i.e., there is no guarantee that all messages arrive. On the other hand, *some* messages will not get lost, or more precisely: if you repeat sending a certain message long enough, it will eventually arrive. This property is called *fairness*. Now the question is whether one can write a protocol (or a program) that satisfies the following two constraints, provided that fairness holds:

- *safety:* at any moment, Y is a prefix of X;
- *liveness:* every x_i will eventually be written as y_i on Y.

Hence, safety expresses that R will only write a correct initial part of the tape X into Y. This is easily achieved by allowing R to never write anything, but the liveness property says that R cannot linger on forever: every bit of X should eventually appear in Y.

In the protocol below, the construct '`send` msg `until` ψ' means that the message msg is repeatedly sent, until the Boolean ψ has become true. The

test $K_R(x_i)$ intuitively means that Receiver knows that the i-th element of X is equal to x_i: more precisely, it is true if Receiver receives a bit b and the last bit he has written is x_{i-1}. The other Booleans, like $K_S K_R(x_i)$ are supposed to express that S has received the message "$K_R(x_i)$".

PROTOCOL FOR S:

$S1$ i :=0
$S2$ while true do
$S3$ begin read x_i;
$S4$ send x_i until $K_S K_R(x_i)$;
$S5$ send "$K_S K_R(x_i)$" until $K_S K_R K_S K_R(x_i)$
$S6$ i := i + 1
$S7$ end

PROTOCOL FOR R:

$R2$ when $K_R(x_0)$ set i :=0
Nec while true do
$R3$ begin write x_i;
$R4$ send "$K_R(x_i)$" until $K_R K_S K_R(x_i)$;
$R5$ send "$K_R K_S K_R(x_i)$" until $K_R(x_{i+1})$
$R6$ i := i + 1
$R7$ end

An important aspect of the protocol is that Sender at line $S5$ does not continue reading X and does not yet add 1 to the counter i. We will show why this is crucial for guaranteeing safety. For, suppose that the lines $S5$ and $R5$ would be absent, and that instead line $R4$ would read as follows:

$R4'$ send "$K_R(x_i)$" until $K_R(x_{i+1})$;

Suppose also, as an example, that $X = \langle a, a, b, \ldots \rangle$. Sender starts by reading x_0, an a, and sends it to R. We know that an instance of that a will arrive at a certain moment, and so by line $R3$ it will be written on Y. Receiver then acts as it should and sends an acknowledgement ($R4'$) that will also arrive eventually, thus Sender continues with $S6$ followed by $S3$: once again it reads an a and sends it to Receiver. The latter will eventually receive an instance of that a, but will not know how to interpret it: "is this symbol a a repetition of the previous one, because Sender does not know that I know what x_0 is, or is this a the next element of the input tape, x_1"? This would clearly endanger safety.

One can show that knowledge of depth four is sufficient and necessary to comply with the specification. As a final remark on the protocol, it be noted

that most programming languages do not refer to epistemic notions. So in general a protocol like the one described above still needs to be transformed into 'a real program'. It is indeed possible to rewrite this protocol without using any knowledge operators. The result is known as the 'alternating bit protocol' (see notes for references). □

We now give an example where some typical reasoning about ignorance, and iterations of everybody's knowledge, are at stake.

Example 2.4 (Consecutive numbers) Two agents, a (Anne) and b (Bill) are facing each other. They see a number on each other's head, and those numbers are consecutive numbers n and $n+1$ for a certain $n \in \mathbb{N}$. They both know this, and they know that they know it, etc. However, they do not have any other a priori knowledge about their own number. Let us assume we have 'atoms' a_n and b_n, for every $n \in \mathbb{N}$, expressing that the number on Anne's head equals n, and that on Bill's head reads n, respectively.

Suppose that in fact a_3 and b_2 are true. Assuming that the agents only see each other's number and that it is common knowledge that the numbers are consecutive, we now have the following (if you find it hard to see why these statements are true, move on and collect some technical tools first, and then try to do Exercise 2.10):

1. $K_a b_2$
   Anne can see b's number.
   For the same reason, we have $K_b a_3$.
2. $K_a(a_1 \vee a_3)$
   expressing that Anne knows that her number is either 1 or 3.
   Similarly, we of course have $K_b(b_2 \vee b_4)$.
3. $K_a K_b(b_0 \vee b_2 \vee b_4)$
   This follows from the previous item.
   Similarity induces that we also have $K_b K_a(a_1 \vee a_3 \vee a_5)$
4. $K_a K_b K_a(b_0 \vee b_2 \vee b_4)$
   If you find it hard to see this, wait until we have explained a formal semantics for this situation, in Section 2.2.2.
   In a similar vein, we have $K_b K_a K_b(a_1 \vee a_3 \vee a_5)$

The above sequence of truths for the real state characterised by $(a_3 \wedge b_2)$ has a number of intriguing consequences for iterated knowledge concerning a and b. This is easier to see if we also enumerate a number of 'epistemic possibilities' for the agents. Let us suppose that the aim of this scenario is, for both Anne and Bill, to find out the number on their head. For this, we introduce $\mathsf{win_a}$ for "Anne knows the number on her head", and similarly, $\mathsf{win_b}$.

5. $\hat{K}_a a_1 \wedge \hat{K}_a a_3$
   Anne considers it possible that her number is 1, but she also considers it possible that she is wearing 3.
   Similarly, we obtain $\hat{K}_b b_2 \wedge \hat{K}_a b_4$

6. $K_a(\neg \mathsf{win_a} \wedge \neg \mathsf{win_b})$
   Note that only if one of the numbers is 0, can somebody know her or his number: anyone who sees 0 knows to have 1. Anne considers two situations possible: one in which $a_1 \wedge b_2$ holds, and another in which $a_3 \wedge b_2$ is true. In neither of those situations is an agent wearing 0, so that Anne knows there is no winner.
   Similarly, we have $K_b(\neg \mathsf{win_a} \wedge \neg \mathsf{win_b})$
7. $E_{\{a,b\}}\neg a_5 \wedge \neg E_{\{a,b\}}E_{\{a,b\}}\neg a_5$
   Anne and Bill know that Anne does not have 5 (because Bill knows she has a 3, and Anne knows she has either 1 or 3). However, not everybody knows that everybody knows this, since we have $\hat{K}_b\hat{K}_a a_5$: Bill thinks it possible that the situation is such that Anne has 3 and Bill 4, in which case Anne would see Bill's 4, and considers it possible that she has 5! Once again, clear semantics may help in verifying these properties: see Section 2.2.2, especially Exercise 2.10.
8. $E_{\{a,b\}}(\neg \mathsf{win_a} \wedge \neg \mathsf{win_b}) \wedge \neg E_{\{a,b\}}E_{\{a,b\}}(\neg \mathsf{win_a} \wedge \neg \mathsf{win_b})$
   Although everybody knows that neither of the agents can win (this follows from item 6), it is not the case that this very fact is known by everybody! To see the latter, note that a thinks it possible that she wears 1 and b 2, in which case b would consider it possible that a's number is 1 and his 0, in which case a would know her number: $\hat{K}_a\hat{K}_b(a_1 \wedge K_a a_1)$. □

It is worthwhile to notice that, in Example 2.4 $\neg E_{\{a,b\}}E_{\{a,b\}}\neg a_5$ holds in the state characterised by $(a_3 \wedge b_2)$ since $\hat{K}_b\hat{K}_a a_5$ is true. We recall a generalised version of this insight in the following exercise, since it will be useful later on.

Exercise 2.5 Let $B = \{b_1, \ldots, b_m\}$ be a group of m agents. Argue that $E_B^n\varphi$ is false if and only if there is a sequence of agents names $a^1, a^2, \ldots a^n$ $(a^i \in B, i \leq n)$ such that $\hat{K}_{a^1}\hat{K}_{a^2}\ldots\hat{K}_{a^n}\neg\varphi$ holds. Note that it is well possible that $n > m$: some agents can reappear in the sequence. □

2.2.2 Semantics

Now let us move to a formal treatment of the logics of knowledge for individual agents within a group. Crucial in the approach to epistemic logic is the use of its *semantics* which uses a special case of Kripke models. In such a model, two notions are of main importance: that of *state* and that of *indistinguishability*. We explain these using a very simple example. We call it the GLO-scenario, named after Groningen, Liverpool, and Otago. Suppose we have one agent, say b, who lives in Groningen. For some reason, he builds a theory about the weather conditions in both Groningen and Liverpool: in Groningen it is either sunny (denoted by the atom g), or not $(\neg g)$. Likewise for Liverpool: sunny (l) or not $(\neg l)$. If for the moment we identify a 'state' with a possible state of the world, then, a priori, we have 4 such states: $\langle g, l\rangle$, in which it is both sunny

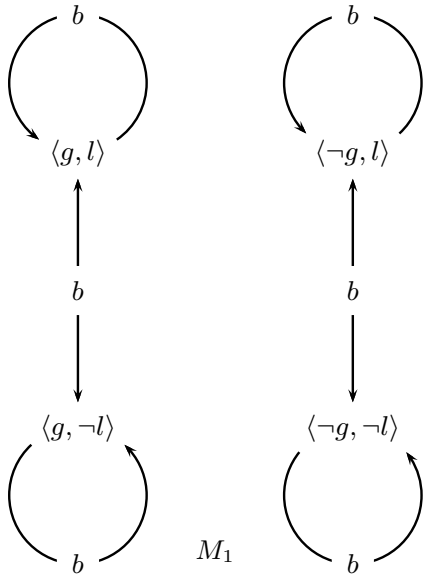

Figure 2.1. Kripke model M_1, representing a GLO scenario.

in Groningen and in Liverpool, $\langle g, \neg l\rangle$ in which the weather in Groningen is again sunny but not in Liverpool, etc. Since b is situated in Groningen, we can assume that he is aware of the weather in Groningen, but not of that in Liverpool. In other words: he cannot distinguish state $\langle g, l\rangle$ from $\langle g, \neg l\rangle$, neither can he tell the difference between $\langle \neg g, l\rangle$ and $\langle \neg g, \neg l\rangle$.

This situation is represented in the Kripke model M_1 of Figure 2.1, where indistinguishability of agent b is represented by an arrow labelled with b. The points in this model are called states, which are in this case indeed states of the world. An arrow labelled with an agent b going from state s to t, is read as: 'given that the state is s, as far as b's information goes, it might as well be t', or 'in state s, agent b considers it possible that the state in fact is t', or, in the case of the models we will mostly consider, 'agent b cannot distinguish state s from state t'. The latter description refers to an equivalence relation (cf. Definition 2.13): no agent is supposed to distinguish s from itself; if t is indistinguishable from s then so is s from t and, finally, if s and t are the same for agent b, and so are t and u, then b cannot tell the difference between s and u. Note that the accessibility relation in model M_1 of Figure 2.1 is indeed an equivalence relation.

Definition 2.6 Given a countable set of atomic propositions P and a finite set of agents A, a *Kripke model* is a structure $M = \langle S, R^A, V^P\rangle$, where

- S is a set of states. The set S is also called the domain $\mathcal{D}(M)$ of M.
- R^A is a function, yielding for every $a \in A$ an accessibility relation $R^A(a) \subseteq S \times S$. We will often write R_a rather than $R^A(a)$ and freely mix a prefix notation $(R_a st)$ with an infix notation $(sR_a t)$.
- $V^P : P \to 2^S$ is a valuation function that for every $p \in P$ yields the set $V^P(p) \subseteq S$ of states in which p is true.

We will often suppress explicit reference to the sets of P and A, and represent a model as $M = \langle S, R, V\rangle$. In such a case, we may also write V_p rather than $V(p)$, for any atom p and the valuation V.

If we know that all the relations R_a in M are equivalence relations, we call M an *epistemic model*. In that case, we write $\sim_a$ rather than R_a, and we represent the model as $M = \langle S, \sim, V\rangle$. □

We will interpret formulas in states. Note that in M_1 of Figure 2.1, the names of the states strongly suggest the valuation V_1 of M_1: we for instance have $\langle g, \neg l\rangle \in V_1(g)$ and $\langle g, \neg l\rangle \notin V_1(l)$. As to the epistemic formulas, in M_1, we want to be able to say that if the state of the world is $\langle g, l\rangle$, i.e., both Groningen and Liverpool show sunny weather, then agent b *knows* it is sunny in Groningen, (since all the states he considers possible, $\langle g, l\rangle$ and $\langle g, \neg l\rangle$ verify this), but he does not know that it is sunny in Liverpool (since he cannot rule out that the real state of the world is $\langle g, \neg l\rangle$, in which case it is not sunny in Liverpool). In short, given the model M_1 and the state $\langle g, l\rangle$, we expect $K_b g \wedge \neg K_b l$ to hold in that state.

Definition 2.7 Epistemic formulas are interpreted on pairs (M, s) consisting of a Kripke model $M = \langle S, R, V\rangle$ and a state $s \in S$. Whenever we write (M, s), we assume that $s \in \mathcal{D}(M)$. Slightly abusing terminology, we will sometimes also refer to (M, s) as a *state*. If M is an epistemic model (see Definition 2.6), (M, s) is also called an *epistemic state*. We will often write M, s rather than (M, s). Such a pair will often be referred to as a *pointed model*.

Now, given a model $M = \langle S, R, V\rangle$ we define that formula φ is true in (M, s), also written as $M, s \models \varphi$, as follows:

$$
\begin{array}{ll}
M, s \models p & \text{iff } s \in V(p) \\
M, s \models (\varphi \wedge \psi) & \text{iff } M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models \neg\varphi & \text{iff not } M, s \models \varphi \\
M, s \models K_a\varphi & \text{iff for all } t \text{ such that } R_a st, M, t \models \varphi
\end{array}
$$

Instead of 'not $M, s \models \varphi$' we also write '$M, s \not\models \varphi$'. The clause for K_a is also phrased as 'K_a is the necessity operator with respect to R_a'. Note that the dual $\hat{K}_a$ obtains the following truth condition, for which it is also dubbed 'a possibility operator with respect to R_a':

$$M, s \models \hat{K}_a\varphi \text{ iff there is a } t \text{ such that } R_a st \text{ and } M, t \models \varphi \qquad (2.1)$$

When $M, s \models \varphi$ for all $s \in \mathcal{D}(M)$, we write $M \models \varphi$ and say that φ is true in M. If $M \models \varphi$ for all models M in a certain class $\mathcal{X}$ (like, for instance, all epistemic models), we say that φ is valid in $\mathcal{X}$ and write $\mathcal{X} \models \varphi$. If $M \models \varphi$ for all Kripke models M, we say that φ is valid, and write $\models \varphi$ or $\mathcal{K} \models \varphi$, where $\mathcal{K}$ is the set of all Kripke models. We use $\not\models$ to deny any of such claims, for instance $M \not\models \varphi$ says that φ is not true in M, meaning that there is some state $s \in \mathcal{D}(M)$ that falsifies it, i.e., for which $M, s \models \neg\varphi$.

If for formula φ there is a state (M, s) such that $(M, s) \models \varphi$, we say that φ is *satisfied* in (M, s), and, if M belongs to a class of models $\mathcal{X}$, we then say that φ is satisfiable in $\mathcal{X}$. □

The keypoint in the truth definition is that an agent a is said to know an assertion φ in a state (M, s) if and only if that assertion is true in all the states he considers possible, given s. Going back to Figure 2.1, we have $M_1, \langle g, l\rangle \models K_b g \wedge \neg K_b l \wedge \neg K_b \neg l$: agent b knows *that* it is sunny in Groningen, but does not know *whether* it is sunny in Liverpool, he does not know that l, nor that $\neg l$.

Kripke semantics makes our epistemic logic *intensional*, in the sense that we give up the property of *extensionality*, which dictates that in any formula, one can always substitute subformulas for different, but equivalent ones. To see that we indeed got rid of extensionality, note that, in Figure 2.1 we have $M_1, \langle g, l\rangle \models (g \wedge l) \wedge (K_b g \wedge \neg K_b l)$ (saying that g and l have the same truth value, but still one can know one without knowing the other).

A second feature of Kripke semantics to note at this point is that it gives us a natural way to interpret arbitrary nested knowledge formulas. For instance, not only does b not know in $\langle g, l\rangle$ that l ($M_1, \langle g, l\rangle \models \neg K_b l$), he knows about his ignorance! (i.e., $M_1, \langle g, l\rangle \models K_b \neg K_b l$.)

Exercise 2.8 M_1 is the model of Figure 2.1.

1. Formalise the following claims:
   a) In state $\langle g, l\rangle$, agent b considers it possible that it is sunny in Groningen, and also that it is sunny in Liverpool, and also that it is not sunny in Liverpool.
   b) In state $\langle \neg g, l\rangle$, agent b knows it is not sunny in Groningen, although he does not know it is sunny in Liverpool.
   c) In state $\langle g, l\rangle$, agent b knows both that he knows that it is sunny in Groningen and that he does not know that it is sunny in Liverpool.
   d) In model M_1, it is true that agent b knows whether it is sunny in Groningen, but he does not know whether it is sunny in Liverpool.
   e) In any model, any agent knows that any fact or its negation holds.
   f) It is not a validity that an agent always knows either a fact, or that he knows its negation.
2. Verify the following:
   a) $M_1, \langle g, l\rangle \models \hat{K}_b g \wedge \hat{K}_b l \wedge \hat{K}_b \neg l$
   b) $M_1, \langle \neg g, l\rangle \models K_b \neg g \wedge \neg K_b l$
   c) $M_1, \langle g, l\rangle \models K_b(K_b g \wedge \neg K_b l)$
   d) $M_1 \models (K_b g \vee K_b \neg g) \wedge (\neg K_b l \wedge \neg K_b \neg l)$
   e) $\models K_a(\varphi \vee \neg\varphi)$
   f) $\not\models K_a \varphi \vee K_a \neg\varphi$ □

We already observed that M_1 of Figure 2.1 is in fact an epistemic model $M_1 = \langle S, \sim, V\rangle$. Since such models will be so prominent in this book, we economise on their representation in figures by representing the equivalence

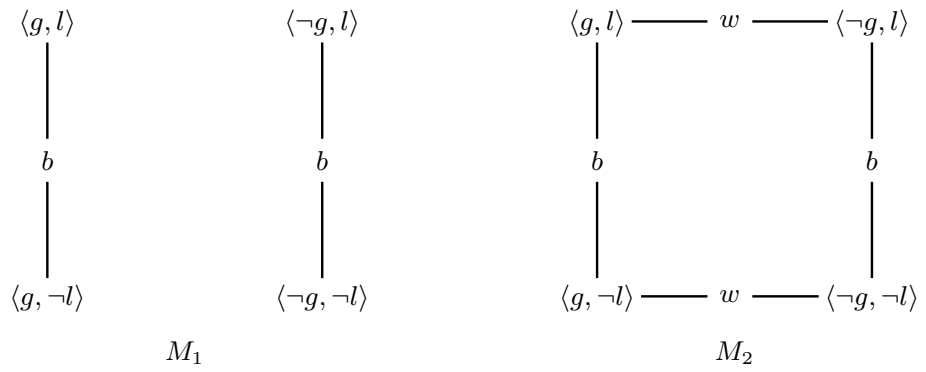

Figure 2.2. Two GLO scenarios.

relations by lines, rather than by arrows. Moreover, we will leave out all reflexive arrows. See model M_1 of Figure 2.2, which is a more economical representation of the model M_1 of Figure 2.1.

One of the ways to make the model M_1 of Figure 2.2 a real multi-agent model is illustrated by the model M_2 in that figure. It represents the situation in which a second agent, w, is situated in Liverpool and knows about the weather there. Note that now we have for instance

$$M_2, \langle g, l\rangle \models K_b g \wedge \neg K_b l \wedge \neg K_w g \wedge K_w l$$

Indeed, we even have $M_2 \models (K_b g \vee K_b \neg g) \wedge (K_w l \vee K_w \neg l)$: whatever the state of the world is, agent b knows whether the sun shines in Groningen, and w knows the weather condition in Liverpool.

But the model M_2 models much more than that: not only does b know the weather conditions in Groningen, and w those in Liverpool, but, apparently, this is also known by both of them! For instance, we can verify (2.2), which says that, in $M_2, \langle g, l\rangle$, agent w does not know whether it is sunny in Groningen, but w *does* know that b knows whether it is sunny there.

$$M_2, \langle g, l\rangle \models \neg K_w g \wedge \neg K_w \neg g \wedge K_w(K_b g \vee K_b \neg g) \tag{2.2}$$

This is verified as follows. First of all, we have $M_2, \langle g, l\rangle \models \neg K_w g$, since, in state $\langle g, l\rangle$, agent w considers $\langle \neg g, l\rangle$ to be the real world, in which g is false, hence he does not know g. More formally, since $\langle g, l\rangle R_w \langle \neg g, l\rangle$ and $M_2, \langle \neg g, l\rangle \models \neg g$, we have $M_2, \langle g, l\rangle \models \neg K_w g$. Given $\langle g, l\rangle$, agent w also considers $\langle g, l\rangle$ itself as a possibility, so not in all the states that w considers possible, $\neg g$ is true. Hence, $M_2, \langle g, l\rangle \models \neg K_w \neg g$. All the states that w considers possible in $\langle g, l\rangle$, are $\langle g, l\rangle$ and $\langle \neg g, l\rangle$. In the first, we have $K_b g$, in the second, $K_b \neg g$. So, in all the states that w considers possible given $\langle g, l\rangle$, we have $K_b g \vee K_b \neg g$. In other words, $M_2, \langle g, l\rangle \models K_w(K_b g \vee K_b \neg g)$.

Exercise 2.9 summarises some multi-agent properties of model M_2: it is important, for the reader to appreciate the remainder of this book, that he or she is sure to be able to do that exercise!

Exercise 2.9 Let M_2 be as above. Verify that:

1. $M_2, \langle g, l\rangle \models g \wedge \hat{K}_w K_b \neg g$
   Although g is true in $\langle g, l\rangle$, agent w considers it possible that b knows $\neg g$!
2. $M_2, \langle g, l\rangle \models (g \wedge l) \wedge \hat{K}_b \hat{K}_w (\neg g \wedge \neg l)$
   Although $(g \wedge l)$ is true at $\langle g, l\rangle$, agent b considers it possible that w considers it possible that $\neg g \wedge \neg l$
3. $M_2 \models K_w(g \rightarrow K_b g)$
   Everywhere in the model, it holds that w knows that, whenever it is sunny in Groningen, b knows it. □

We end the *GLO* story with two remarks regarding 'states' and valuations. In the two models we have considered so far (M_1 and M_2), the notion of epistemic state and valuation coincided. In general, this need not be the case. First of all, not every possible valuation (a possible state of the world) needs to be present in a(n) (epistemic) model. Suppose that we enrich our language with an atom o, denoting whether it is sunny in Otago, New Zealand. Then, obviously, the valuation that would denote the state in which $g \wedge l \wedge o$ is true for instance would not appear in any model, since a background theory, based on geographical insights, ensures that o can never be true at the same time with g or with l. Secondly, in general it is possible that one and the same valuation, or one and the same possible world, can represent different epistemic states.

Let us, rather than enrich the set of propositions in our *GLO* example, enrich the set of agents with an agent h who, as a matter of fact, happens to live in Otago. Suppose the world is in such a state that g and l are true. Moreover, we make the same assumptions about b and w as we did earlier: b knows about g, and w knows about l, and all this is known by each agent. See model M_2 of Figure 2.3, which is exactly the same epistemic model as M_2 of Figure 2.2.

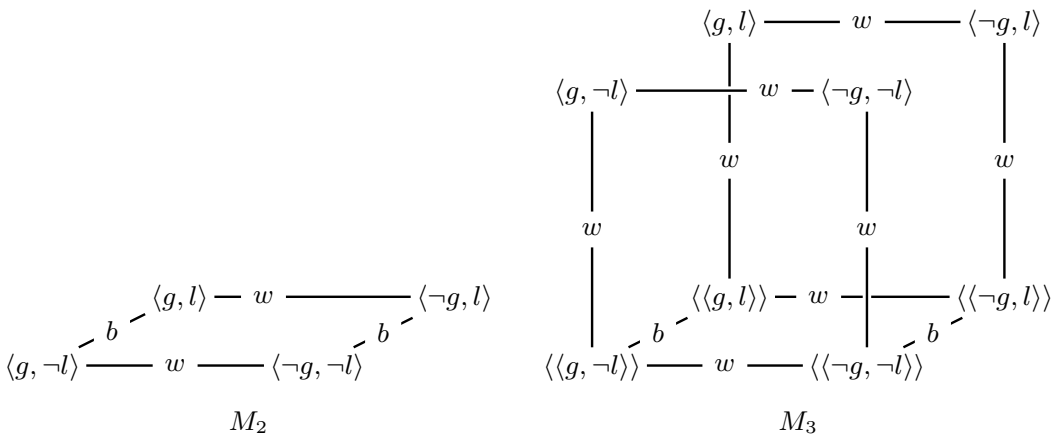

Figure 2.3. GLO scenarios, one with multiple occurrences of the same valuation.

This book is about the dynamics of epistemics, of which a first example will now be given. It is night-time in Otago, but since the three agents are involved in writing a book, this fact does not stop h from calling his European co-authors. He first calls w. They have some social talk (w tells h that the sun is shining in Liverpool) and they discuss matters of the book. Then h tells w truthfully what he is going to do next: call b and report what h and w have discussed so far. The poor w does not have a clue whether this also means that h will reveal the weather condition of Liverpool (l) to b. Hence, w considers two epistemic states possible that have the same valuation: in one of them, we have $g \wedge l \wedge \neg K_b l$ and in the other, $g \wedge l \wedge K_b l$. The first state corresponds with the state $\langle\langle g, l\rangle\rangle$ in model M_3 of Figure 2.3, and which is exactly at the position $\langle g, l\rangle$ of model M_2, the second with $\langle g, l\rangle$ 'just above' the previous one in the figure. In model M_2 every two states $\langle x, y\rangle$ and $\langle\langle x, y\rangle\rangle$ have the same valuation, although they are not the same states! Given the fact that one in general has different states with the same valuation, it is not a good idea to name a state after its valuation, as we have done in M_1 and M_2. As an aside, we have not modelled h's knowledge in model M_3.

Note, in passing, that we applied another economic principle in representing epistemic models, in M_3 of Figure 2.3: although, by transitivity of such models, there should be a connecting line between $\langle g, l\rangle$ and *both* occurrences of $\langle \neg g, l\rangle$, we leave out one of them in our drawing, relying on the reader's ability to 'complete the transitive closure' of our representation.

What exactly does the epistemic state $M_3, \langle\langle g, l\rangle\rangle$ describe? Well, it is the situation in which h did *not* tell b about l (since b still considers a state possible in which $\neg l$), but in which both b and w know that h might have informed b about the truth value of l. Note for instance that we have that b knows that w does not know whether b knows l: we have

$$M_3, \langle\langle g, l\rangle\rangle \models l \wedge \neg K_b l \wedge K_b(\neg K_w K_b l \wedge \neg K_w \neg K_b l) \tag{2.3}$$

(2.3) expresses that, although l holds, b does not know this, but b knows that w neither knows that b knows that l, nor that b does not know that l.

Exercise 2.10 (Modelling consecutive numbers) Read again the description of the Consecutive Numbers example, Example 2.4. Check the truth definition for E_B at Definition 2.30.

1. Draw the appropriate model for it. Omit reflexive arrows. Call the model M and denote the factual state as $\langle 3, 2\rangle$.
2. Show that $M, \langle 1, 0\rangle \models K_a a_1 \wedge \mathsf{win_a}$
3. What are the states in which an agent can win the game? What is the difference between them, in terms of epistemic properties, given that the real state is $\langle 3, 2\rangle$?
4. Verify the statements 1–8 of Example 2.4 in state $M, \langle 3, 2\rangle$. □

How appropriate are Kripke models to represent knowledge? A possible answer has it that they incorporate too strong properties. This is sometimes referred to as *logical omniscience*; it is technically summarised in the following proposition.

Proposition 2.11 Let φ, ψ be formulas in $\mathcal{L}_K$, and let K_a be an epistemic operator for $a \in A$. Let $\mathcal{K}$ be the set of all Kripke models, and $\mathcal{S}5$ the set of Kripke models in which the accessibility relation is an equivalence (see Definition 2.13). Then the following hold:

- $\mathcal{K} \models K_a\varphi \wedge K_a(\varphi \rightarrow \psi) \rightarrow K_a\psi$ *LO*1
- $\mathcal{K} \models \varphi \Rightarrow \models K_a\varphi$ *LO*2
- $\mathcal{K} \models \varphi \rightarrow \psi \Rightarrow \models K_a\varphi \rightarrow K_a\psi$ *LO*3
- $\mathcal{K} \models \varphi \leftrightarrow \psi \Rightarrow \models K_a\varphi \leftrightarrow K_a\psi$ *LO*4
- $\mathcal{K} \models (K_a\varphi \wedge K_a\psi) \rightarrow K_a(\varphi \wedge \psi)$ *LO*5
- $\mathcal{K} \models K_a\varphi \rightarrow K_a(\varphi \vee \psi)$ *LO*6
- $\mathcal{S}5 \models \neg(K_a\varphi \wedge K_a\neg\varphi)$ *LO*7

□

The fact that the above properties hold in all Kripke models is referred to as the problem of *logical omniscience* since they express that agents are omniscient, perfect logical reasoners. For example, *LO*1 says that knowledge is closed under consequences. *LO*2 expresses that agents know all ($\mathcal{S}5$-)validities. *LO*3-*LO*6 all assume that the agent is able to make logical deductions with respect to his knowledge, and, on top of that, *LO*7 ensures that his knowledge is internally consistent. The properties of Proposition 2.11 reflect *idealised* notions of knowledge, that do not necessarily hold for human beings. For example, many people do not know all tautologies of propositional logic, so *LO*2 does not hold for them. We will see how, in many systems, the properties of Proposition 2.11 are nevertheless acceptable. If the properties mentioned here are unacceptable for a certain application, a possible world approach to knowledge is probably not the best option: all listed properties are valid in all Kripke models, except for *LO*7, which is only true on *serial* models (see Definition 2.13).

The popularity of Kripke semantics reaches far further than the area of epistemics. Depending on the intended interpretation of the modal operator, one can freely write, rather than K_a, other symbols for these operators. Interpretations that are very common (and hence are supposed to satisfy the properties of Proposition 2.11) are 'φ is believed' $B_a\varphi$, or 'φ is always the case' ($\Box\varphi$), 'φ is a desire' ($D_a\varphi$), 'φ is obligatory' ($\bigcirc\varphi$), 'φ is provable' ($\Box\varphi$), or 'φ is a result of executing program π' ($[\pi]\varphi$).

All this does not mean that these notions have exactly the same logic, or properties. Rather, opting to model a specific operator using the Kripke semantics of Definition 2.7, Proposition 2.11 gives some minimal properties of it. One of the main features of Kripke semantics is that one can, in a modular fashion, put additional constraints on the accessibility relation, to obtain some extra modal validities. As a simple example, if in the model M the accessibility relation R_a is reflexive (i.e., $\forall s R_a s s$), then M satisfies $K_a\varphi \rightarrow \varphi$ (see Exercise 2.12). Hence, if $\mathcal{T}$ is the class of all reflexive models, i.e., $\mathcal{T} = \{M = \langle S, R, V\rangle \mid \text{ every } R_a \text{ is reflexive}\}$, then $\mathcal{T} \models K_a\varphi \rightarrow \varphi$.

Exercise 2.12 Show, that if $M = \langle S, R, V\rangle$ is such that R_a is reflexive, then M satisfies the truth axiom: $M \models K_a\varphi \rightarrow \varphi$. □

We now define some classes of models that are of crucial importance in epistemic logic.

Definition 2.13 Recall that R stands for a family of accessibility relations, one R_a for each $a \in A$.

1. The class of all Kripke models is sometimes denoted $\mathcal{K}$. Hence, $\mathcal{K} \models \varphi$ coincides with $\models \varphi$.
2. R_a is said to be *serial* if for all s there is a t such that R_ast
   The class of serial Kripke models $\{M = \langle S, R, V\rangle \mid$ every R_a is serial$\}$ is denoted by $\mathcal{KD}$.
3. R_a is said to be *reflexive* if for all s, R_ass
   The class of reflexive Kripke models $\{M = \langle S, R, V\rangle \mid$ every R_a is reflexive$\}$ is denoted by $\mathcal{T}$.
4. R_a is *transitive* if for all s, t, u, if R_ast and R_atu then R_asu
   The class of transitive Kripke models is denoted by $\mathcal{K}4$.
   The class of reflexive transitive models is denoted by $\mathcal{S}4$.
5. R_a is *Euclidean* if for all s, t, and u, if R_ast and R_asu then R_atu
   The class of transitive Euclidean models is denoted by $\mathcal{K}45$.
   The class of serial transitive Euclidean models is denoted by $\mathcal{KD}45$.
6. R_a is an *equivalence relation* if R_a is reflexive, transitive, and symmetric (for all s, t, if R_ast then R_ats). Equivalently, R_a is an equivalence relation if R_a is reflexive, transitive and Euclidean.
   The class of Kripke models with equivalence relations is denoted by $\mathcal{S}5$.□

These classes of models will be motivated shortly, but, as said before, the main emphasis in this book will be on $\mathcal{S}5$.

How much can two epistemic states M, s and M', s' differ without affecting the knowledge of any agent? In other words, how expressive is our epistemic language $\mathcal{L}_K$, to which granularity can it distinguish models from each other? We address the issue of expressivity of several languages introduced in this book in Chapter 8, but now present a basic underlying notion and result.

Definition 2.14 (Bisimulation) Let two models $M = \langle S, R, V\rangle$ and $M' = \langle S', R', V'\rangle$ be given. A non-empty relation $\mathfrak{R} \subseteq S \times S'$ is a bisimulation iff for all $s \in S$ and $s' \in S'$ with $(s, s') \in \mathfrak{R}$:

atoms $s \in V(p)$ iff $s' \in V'(p)$ for all $p \in P$
forth for all $a \in A$ and all $t \in S$, if $(s, t) \in R_a$, then there is a $t' \in S'$ such that $(s', t') \in R'_a$ and $(t, t') \in \mathfrak{R}$
back for all $a \in A$ and all $t' \in S'$, if $(s', t') \in R'_a$, then there is a $t \in S$ such that $(s, t) \in R_a$ and $(t, t') \in \mathfrak{R}$

We write $(M, s) \leftrightarrows (M', s')$, iff there is a bisimulation between M and M' linking s and s'. Then we call (M, s) and (M', s') bisimilar. □

Obviously, for M, s and M', s' to bisimulate each other, the **atoms**-clause guarantees that there is agreement on objective formulas (not only between s and s', but recursively in all states that can be reached), the **forth**-clause preserves ignorance formulas going from M, s to M', s', and the **back**-clause preserves knowledge. This is made more precise in the (proof of) the following theorem, which says that the epistemic language $\mathcal{L}_K$ cannot distinguish bisimilar models. We write $(M, s) \equiv_{\mathcal{L}_K} (M', s')$ if and only if $(M, s) \models \varphi$ iff $(M', s') \models \varphi$ for all formulas $\varphi \in \mathcal{L}_K$.

Theorem 2.15 For all pointed models (M, s) and (M', s'), if $(M, s) \leftrightarrows (M', s')$, then $(M, s) \equiv_{\mathcal{L}_K} (M', s')$. □

Proof We proceed by induction on φ.

Base case Suppose $(M, s) \leftrightarrows (M', s')$. By **atoms**, it must be the case that $(M, s) \models p$ if and only if $(M', s') \models p$ for all $p \in P$

Induction hypothesis For all pointed models (M, s) and (M', s'), if we have that $(M, s) \leftrightarrows (M', s')$, then it follows that $(M, s) \models \varphi$ if and only if $(M', s') \models \varphi$.

Induction step

negation Suppose that $(M, s) \models \neg\varphi$. By the semantics this is the case if and only if $(M, s) \not\models \varphi$. By the induction hypothesis this is equivalent to $(M', s') \not\models \varphi$, which is the case if and only if $(M', s') \models \neg\varphi$.

conjunction Suppose that $(M, s) \models \varphi_1 \wedge \varphi_2$, with, by the induction hypothesis, the theorem proven for φ_1 and φ_2. By the semantics the conjunction $\varphi_1 \wedge \varphi_2$ is true in (M, s) if and only if $(M, s) \models \varphi_1$ and $(M, s) \models \varphi_2$. By the induction hypothesis this is equivalent to $(M', s') \models \varphi_1$ and $(M', s') \models \varphi_2$. By the semantics this is the case if and only if $(M', s') \models \varphi_1 \wedge \varphi_2$.

individual epistemic operator Suppose $(M, s) \models K_a\varphi$. Take an arbitrary t' such that $(s', t') \in R'_a$. By **back** there is a $t \in S$ such that $(s, t) \in R_a$ and $(t, t') \in \mathfrak{R}$. Therefore, by the induction hypothesis $(M, t) \models \varphi$ if and only if $(M', t') \models \varphi$. Since $(M, s) \models K_a\varphi$, by the semantics $(M, t) \models \varphi$. Therefore $(M', t') \models \varphi$. Given that t' was arbitrary, $(M', t') \models \varphi$ for all t' such that $(s', t') \in R'_a$. Therefore by the semantics $(M', s') \models K_a\varphi$.

The other way around is analogous, but then **forth** is used. □

So, having a bisimulation is *sufficient* for two states to verify the same formulas, and hence to represent the same knowledge. In Chapter 8 we will see that it is not *necessary*, however. Note that the proof given above did not depend on R or R' being reflexive, transitive, or Euclidean. So indeed it also holds for other modal logics than *S5*.

2.2.3 Axiomatisation

A logic is a set of formulas. One way to characterise them is semantically: take a set of formulas that is valid in a class of models. An *axiomatisation*

all instantiations of propositional tautologies	
$K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$	distribution of K_a over $\to$
From φ and $\varphi \to \psi$, infer ψ	modus ponens
From φ, infer $K_a\varphi$	necessitation of K_a

Table 2.1. The basic modal system **K**.

is a syntactic way to specify a logic: it tries to give a core set of formulas (the axioms) and inference rules, from which all other formulas in the logic are derivable. The first axiomatisation **K** we present gives the axioms of a minimal modal logic: it happens to capture the validities of the semantic class $\mathcal{K}$, i.e., the class of *all* Kripke models. We assume that the modal operators K_a that appear in any axiomatisation are ranging over a set of agents A, which we will not always explicitly mention.

Definition 2.16 The basic epistemic logic **K**, where we have an operator K_a for every $a \in A$, is comprised of all instances of propositional tautologies, the K axiom, and the derivation rules Modus Ponens (MP) and Necessitation (Nec), as given in Table 2.1. □

These axioms and rules define the weakest multi-modal logic **K**, for a set of agents A. The distribution axiom is sometimes also called the axiom K. In formal proofs, we will sometimes refer to the first axiom as $Prop$, and to the two inference rules as MP and Nec, respectively. A modal operator satisfying axiom K and inference rule necessitation is called a *normal* modal operator. All systems that we study in this book will be extensions of **K**. Note that the φ in the first axiom about instantiations of propositional tautologies does not have to be in the propositional *language*: examples of formulas that we obtain by $Prop$ are not only $p \vee \neg p$ and $p \to (q \to p)$ but also $K_a\hat{K}_b\neg q \vee \neg K_a\hat{K}_b\neg q$ and $K_ap \to (K_b(p \vee \hat{K}_bp) \to K_ap)$. Also note that the necessitation rule does not tell us that φ *implies* $K_a\varphi$: it merely states that for any theorem φ that can be derived in this system, we get another one for free, i.e., $K_a\varphi$. Admittedly, we have not yet made precise what it means to be a theorem of an axiomatisation.

Definition 2.17 Let **X** be an arbitrary axiomatisation with axioms Ax_1, Ax_2, …, Ax_n and rules $Ru_1, Ru_2, \ldots Ru_k$, where each rule $Ru_j (j \leq k)$ is of the form "From $\varphi_1, \ldots \varphi_{j_{ar}}$ infer φ_j". We call j_{ar} the *arity* of the rule. Then, a *derivation* for φ within **X** is a finite sequence $\varphi_1, \ldots \varphi_m$ of formulas such that:

1. $\varphi_m = \varphi$;
2. every φ_i in the sequence is
   a) either an instance of one of the axioms $Ax_1, Ax_2, \ldots Ax_n$
   b) or else the result of the application of one of the rules $Ru_j (j \leq k)$ to j_{ar} formulas in the sequence that appear before φ_i.

If there is a derivation for φ in **X** we write $\vdash_{\mathbf{X}} \varphi$, or, if the system **X** is clear form the context, we just write $\vdash \varphi$. We then also say that φ is a *theorem* of **X**, or that **X** proves φ. □

In Table 2.1 and Definition 2.17 we sloppily used the terms *axiom* and *formula* interchangeably, or, more precisely, we use the same meta-variables $\varphi, \psi, \ldots$, for both. Strictly speaking, a formula should not contain such variables. For instance, the *formula* $K_a(p \to q) \to (K_a p \to K_a q)$ is an instance, and hence, derivable in one step, from the axiom *scheme* K. This sloppiness is almost always harmless. In fact, it makes it easy to for instance define that two axioms A and B are *equivalent* with respect to an axiomatisation $\mathbf{X}$. For any system $\mathbf{X}$ and axiom φ, let $\mathbf{X}+\varphi$ denote the axiom system that has axiom φ added to those of $\mathbf{X}$, and the same rules as $\mathbf{X}$. Moreover, let $\mathbf{X} - \varphi$ denote the axiom system $\mathbf{X}$, but without the axiom φ. Then, two axioms A and B are equivalent with respect to $\mathbf{X}$ if $\vdash_{(\mathbf{X}-A)+B} A$ and $\vdash_{(\mathbf{X}-B)+A} B$.

The following exercise establishes two simple facts about this minimal modal logic $\mathbf{K}$:

Exercise 2.18 Let $\vdash \varphi$ stand for $\vdash_{\mathbf{K}} \varphi$.

1. Show that the rule of *Hypothetical Syllogism* (HS) is derivable:
   $\vdash \varphi \to \chi, \vdash \chi \to \psi \Rightarrow \ \vdash \varphi \to \psi$
2. Show $\vdash \varphi \to \psi \Rightarrow \vdash K_a\varphi \to K_a\psi$.
3. Show that, given $\mathbf{K}$, the distribution axiom is equivalent to
   K' $(K_a\varphi \wedge K_a(\varphi \to \psi)) \to K_a\psi)$
   and also to
   K'' $K_a(\varphi \wedge \psi) \to (K_a\varphi \wedge K_a\psi)$
4. Show that $\vdash (K_a\varphi \wedge K_a\psi) \to K_a(\varphi \wedge \psi)$. □

The properties of logical omniscience $LO1 - LO6$ that we gave in Proposition 2.11 are also derivable in $\mathbf{K}$. Although these properties indicate that the notion of knowledge that we are formalising is quite strong and overidealised, at the same time there appear to be other properties of knowledge that are often desirable or natural, and which are not theorems in $\mathbf{K}$, like $K_a\varphi \to \varphi$: what is known, must be true. We will follow standard practice in computer science and embrace (a number of) the additional axioms given in Table 2.2.

The truth-axiom will also be referred to as axiom T and expresses that knowledge is *veridical*: whatever one claims to know, must be true. In other words, it not only inconsistent to say 'I know it is Tuesday, although in fact it is Wednesday', but this also applies when referring to knowledge of a third person, i.e., it makes no sense to claim 'although Bob knows that Ann holds an Ace of spades, it is in fact a Queen of hearts'.

The other two axioms specify so-called *introspective agents*: an agent not only knows what he knows (positive introspection), but also, he knows what he does not know (negative introspection). These axioms will also be denoted

$K_a\varphi \to \varphi$	truth
$K_a\varphi \to K_aK_a\varphi$	positive introspection
$\neg K_a\varphi \to K_a\neg K_a\varphi$	negative introspection

Table 2.2. Axioms for knowledge.

by axiom 4 and axiom 5, respectively. For human agents especially 5 seems an unrealistically strong assumption, but for artificial agents, negative introspection often makes sense. In particular, when in the underlying semantics access for agents is interpreted as indistinguishability, both types of introspection come for free.

A special case of this is provided by the so-called *interpreted systems*, one of the main paradigms for epistemic logic in computer science. Here, the idea is that every agent, or processor, has a *local view* of the system, which is characterised by the value of the local variables. A process a cannot distinguish two global states s and t, if the assignment to a's values are the same, in s and t. One easily verifies that under this definition of access for agent a, knowledge verifies all properties of Table 2.2.

We can now define some of the main axiomatic systems in this book. Recall that $\mathbf{X} + \varphi$ denotes the axiom system that has axiom φ added to those of $\mathbf{X}$, and the same rules as $\mathbf{X}$.

Definition 2.19 We define the following axiom systems: see Figure 2.4. □

Exercise 2.20 Consider the following axiom B: $\varphi \to K_a \hat{K}_a \varphi$.
Show that axiom 5 and B are equivalent with respect to $\mathbf{K} + T + 4$, i.e., show that $\vdash_{\mathbf{K}+T+4+5} B$ and $\vdash_{\mathbf{K}+T+4+B} 5$. □

Although the focus in this book is on **S5**, we end this section by showing a nice kind of 'modularity' in adding axioms to the minimal modal logic **K**. Recall that a logic is a set of formulas, and we have now seen two ways to characterise a logic: as a set of *validities* of a class of models, and as a set of *derivables* of an axiom system. The following theorem is folklore in modal logic: for completeness of **S5** see also Theorem 7.7 of Chapter 7.

Theorem 2.21

1. (Soundness and completeness)
   Axiom system **K** is sound and complete with respect to the semantic class $\mathcal{K}$, i.e., for every formula φ, we have $\vdash_{\mathbf{K}} \varphi$ iff $\mathcal{K} \models \varphi$.
   The same holds for **T** w.r.t. $\mathcal{T}$, for **S4** w.t.t. $\mathcal{S}4$ and, finally, for **S5** w.r.t. $\mathcal{S}5$.
2. (Finite models and decidability)
   Each of the systems mentioned above has the finite model property: any φ is satisfiable in a class $\mathcal{X}$ if and only if it is satisfiable in a finite model of that class.

$\mathbf{T} = \mathbf{K} + T$
$\mathbf{S4} = \mathbf{T} + 4$
$\mathbf{S5} = \mathbf{S4} + 5$

Figure 2.4. Some basic axiom systems.

Moreover, all the systems mentioned are *decidable*: for any class $\mathcal{X}$ mentioned, there exists a decision procedure that determines, in a finite amount of time, for any φ, whether it is satisfiable in $\mathcal{X}$ or not. □

We also define a notion of *derivability from premises*, which gives rise to *strong completeness*.

Definition 2.22 Let $\Box$ be an arbitrary modal operator. An inference rule *Ru* is called a *necessitation rule for* $\Box$ if it is of the form "From φ, infer $\Box\varphi$". Let again $\mathbf{X}$ be an arbitrary axiomatisation with axioms Ax_1, Ax_2, ..., Ax_n and rules $Ru_1, Ru_2, \ldots Ru_k$, where each rule $Ru_j (j \leq k)$ is of the form "From $\varphi_1, \ldots \varphi_{j_{ar}}$ infer φ_j". Define the *closure under necessitation rules* of $\mathbf{X}$ as the smallest set $Cl_{Nec}(\mathbf{X}) \supseteq \{Ax_1, \ldots, Ax_n\}$ such that for any $\psi \in Cl_{Nec}(\mathbf{X})$, and necessitation rule for $\Box$, also $\Box\psi \in Cl_{Nec}(\mathbf{X})$. Let $\Gamma \cup \{\varphi\}$ be a set of formulas. A *derivation for* φ *from* Γ is a finite sequence $\varphi_1, \ldots \varphi_m$ of formulas such that:

1. $\varphi_m = \varphi$;
2. every φ_i in the sequence is
   a) either an instance of one of the schemes in $Cl_{Nec}(\mathbf{X})$
   b) or a member of Γ
   c) or else the result of the application of one of the rules $Ru_j (j \leq k)$ which is not a necessitation rules to j_{ar} formulas in the sequence that appear before φ_i.

If there is a derivation from Γ for φ in $\mathbf{X}$ we write $\Gamma \vdash_{\mathbf{X}} \varphi$, or, if the system $\mathbf{X}$ is clear form the context, we just write $\Gamma \vdash \varphi$. We then also say that φ is *derivable* in $\mathbf{X}$ from the premises Γ.

Given a class of models $\mathcal{C}$, we say that $\mathbf{X}$ is *strongly complete* with respect to $\mathbf{X}$, if for any Γ and φ, we have

$$\Gamma \vdash \varphi \text{ only if (for all } M \in \mathcal{C}, s \in M : \ M, s \models \Gamma \text{ implies } M, s \models \varphi)$$

If the 'only if' is replaced by 'if', we say that $\mathbf{X}$ is *strongly sound* with respect to $\mathcal{C}$. □

So, $\Gamma \vdash_{\mathbf{X}} \varphi$ holds, if there is a proof of φ using the premises in Γ, but without applying necessitation to them. This constraint guarantees that premises are 'local', or 'private', i.e., not necessarily known to everyone. For instance, without this constraint, we would have $\{K_a p, \neg K_b p\} \vdash_{\mathbf{S5}} \bot$, since allowing necessitation to the first premise would yield $K_b K_a p$ $(*)$. Then, a necessitation step of axiom T gives $K_b(K_a p \to p)$, which, together with $(*)$ and the distribution of K_b over $\to$ gives $K_b p$, which is inconsistent with the second premise.

Theorem 2.23 Axiom system $\mathbf{K}$ is strongly sound and strongly complete with respect to the semantic class $\mathcal{K}$. The same holds for $\mathbf{T}$ w.r.t. $\mathcal{T}$, for $\mathbf{S4}$ w.t.t. $\mathcal{S}4$ and, finally, for $\mathbf{S5}$ w.r.t. $\mathcal{S}5$. □

2.3 Group Notions of Knowledge

The notion of 'everybody knows' (or, general knowledge, as it is sometimes called) was already defined in the previous section (page 12). In this section, we introduce some other notions of group knowledge for multiple agent systems. The main emphasis will be on *common knowledge*, a notion that is important throughout this book.

2.3.1 Language

If one adds the definition of general knowledge to **S5**, the prominent epistemic logic, it is easy to see that this notion of everybody knowing inherits veridicality from K, but, if there is more than one agent in A, the same is not true for the introspection properties. And, indeed, lack of positive introspection for the whole group makes sense: if the agents b and w both hear on the radio that it is sunny in Otago (o), we have $E_{\{b,w\}}o$, but not necessarily $E_{\{b,w\}}E_{\{b,w\}}o$: agent b cannot just assume that w heard this announcement as well (see also Example 2.4, item 7). For a similar reason, negative introspection does not (and should not) automatically carry over to E-knowledge: if w missed out on the radio programme announcing o, we have $\neg E_{\{b,w\}}o$, but how could b infer this? This would be needed to conclude $E_{\{b,w\}}\neg E_{\{b,w\}}o$.

Hence, although it is easy to prove in **S5** that for every $n \geq 1$, $K_a^n\varphi$ is equivalent to $K_a\varphi$, for E_B-knowledge, all iterations $E_B^m\varphi$ and E_B^n are in principle different ($m \neq n$). A limiting, and as we shall see, intriguing notion here is *Common Knowledge*, which intuitively captures the infinite conjunction

$$C_B\varphi = \bigwedge_{n=0}^{\infty} E_B^n\varphi$$

The logic of knowledge with common knowledge is denoted $S5C$. Common knowledge is a very strong notion, and hence can in general only be obtained for *weak* formulas φ (sometimes $C_B\varphi$ is dubbed 'any fool knows φ'). One can intuitively grasp the fact that the number of iterations of the E-operator makes a real difference in practice.

Example 2.24 (Saint Nicholas) Suppose that p stands for "Saint Nicholas does not exist" (on December 5, according to tradition in several countries, Saint Nicholas is supposed to visit homes and to bring presents. Children generally start to disbelieve in his existence when they are around six years old, but for various reasons many children like to pretend to believe in him a little longer. Of course, nobody is supposed to reveal the secret on the family evening itself).

Let F be the family gathering together on this evening, and let a, b and c represent different members of F. Imagine how the family's celebration of Saint Nicholas' Eve would look like if $K_ap \wedge \neg E_Fp$ holds (in which case a will

absolutely not make any comment suggesting the visitor in his Saint's dressing and the white beard is in fact the family's neighbour– note that $\neg E_F p$ implies $\hat{K}_a \neg E_F p$, if knowledge is veridical).

Now compare this to the situation where $E_F p \wedge \neg E_F E_F p$ holds (again, if $\neg K_a K_b p$, family member a will not make any public comment upon Saint Nicholas' non-existence, even though everybody knows the saint does not exist), or an evening in which $E_F E_F p \wedge \neg E_F E_F E_F p$ holds. In the latter case, we might for instance have $\neg K_a K_b K_a p$. In that case, a might without danger reveal his disbelief: since a knows that everybody knows p already, he might wish not to look childish by demonstrating to b that a indeed also belongs to the group of adults 'who know'. However, a might also opt to try and exploit $\hat{K}_a \hat{K}_b \hat{K}_a \neg p$, and challenge b not to reveal to a the wisdom that Saint Nicholas does not exist. Similarly, in case that $E_F E_F p \wedge \neg K_a K_b K_c p$, member a might try to resolve possible complications by informing b that $K_c p$, however a might instead choose to exploit the possible situation that $\hat{K}_b \hat{K}_c \neg p$, and try to bring b in a complex situation in which b has to make an effort not to reveal the secret p to c. This would imply some possible entertainment for a (since $K_a K_c p$) which will in fact not occur (since $K_b K_c p$). □

Another way to grasp the notion of common knowledge is to realise in which situations $C\varphi$ does *not* hold for a group. This is the case as long as someone, on the grounds of their knowledge, considers it a possibility that someone considers it a possibility that someone ... that φ does not hold (see also Exercise 2.37 and the remark above it). The following example illustrates such a situation.

Example 2.25 (Alco at the conference) Alco is one of a group B of visitors at a conference in Barcelona, where at a certain point during the afternoon he becomes bored and decides, in fact as the only member of B, to lounge in the hotel bar. While he is enjoying himself there, an important practical announcement φ is made in the lecture room. Of course at that moment $C_B \varphi$ does not hold, nor even $E_B \varphi$. But now suppose that in the bar the announcement comes through by way of an intercom connected to the lecture room. Then we do have $E_B \varphi$, but not $C_B \varphi$; after all, the other visitors of the conference do not know that Alco knows φ.

After hearing φ, Alco leaves the hotel for some sightseeing in the city. At that moment someone in the lecture room worriedly asks whether Alco knows φ, upon which the programme chair reassures her, and thereby anybody else present in the conference room, that this is indeed the case, because of the intercom. Of course at that moment, $C_B \varphi$ still does not hold! □

Now we are ready to give the definition of the full language of epistemic logic, including common knowledge.

Definition 2.26 (Language $\mathcal{L}_{KC}$ with common knowledge) Let P be a set of atomic propositions, and A a set of agent-symbols. We use $a, b, c, \ldots$ as variables over A, and B as a variable over coalitions of agents, i.e., subsets

of A. The language $\mathcal{L}_{KC}$, the language for multi-agent epistemic logic with common knowledge, is generated by the following BNF:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid C_B\varphi$$

For 'small' groups of agents B, we will sometimes write $C_a\varphi, C_{ab}\varphi$, $C_{abc}\varphi$, ..., rather than $C_{\{a\}}\varphi, C_{\{a,b\}}\varphi, C_{\{a,b,c\}}\varphi, \ldots$. Similarly for general knowledge, $E_B\varphi$. □

Thus, $\mathcal{L}_{KC}$ extends $\mathcal{L}_K$ with a notion of common knowledge, for every group.

Example 2.27 (Byzantine generals) Imagine two allied generals, a and b, standing on two mountain summits, with their enemy in the valley between them[1]. It is generally known that a and b together can easily defeat the enemy, but if only one of them attacks, he will certainly lose the battle.

General a sends a messenger to b with the message m (= "I propose that we attack on the first day of the next month at 8 PM sharp"). It is not guaranteed, however, that the messenger will arrive. Suppose that the messenger does reach the other summit and delivers the message to b. Then K_bm holds, and even K_bK_am. Will it be a good idea to attack? Certainly not, because a wants to know for certain that b will attack as well, and he does not know that yet. Thus, b sends the messenger back with an 'okay' message. Suppose the messenger survives again. Then $K_aK_bK_am$ holds. Will the generals attack now? Definitely not, because b does not know whether his 'okay' has arrived, so $K_bK_aK_bm$ does not hold, and common knowledge of m has not yet been established.

In general, for every $n \geq 0$, one can show the following by induction. Recall that $(K_aK_b)^n$ is the obvious abbreviation for $2n$ knowledge operators K_a and K_b in alternation, starting with K_a.

odd rounds After the messenger has safely brought $2n + 1$ such messages (mostly acknowledgements), $K_b(K_aK_b)^nm$ is true, but $(K_aK_b)^{n+1}m$ is not.

even rounds After the messenger has safely brought $2n + 2$ such messages, one can show the following: $(K_aK_b)^{n+1}m$ is true, but $K_b(K_aK_b)^{n+1}m$ is not.

Thus, common knowledge will never be established in this way, using a messenger. Moreover one can prove that in order to start a coordinated attack, common knowledge of m *is* necessary. □

[1] Maybe this example from the theoretical computer scientists' folklore is not politically very correct, but one can imagine more peaceful variants in which synchronisation is of vital importance, e.g., two robots that have to carry a heavy container together.

Before we move on to semantics, let us spend one paragraph on another prominent notion of group knowledge, a notion that will not play an important role in this book, though. *Implicit* or *distributed knowledge* also helps to understand processes within a group of people or collaborating agents. Distributed knowledge is the knowledge that is implicitly present in a group, and which could become explicit if someone would pull all their knowledge together. For instance, it is possible that no agent knows the assertion ψ, while at the same time the distributed knowledge $D_{ab}\psi$ may be derived from $K_a\varphi \wedge K_b(\varphi \rightarrow \psi)$. An example of distributed knowledge in a group is, for instance, the fact whether two members of that group have the same birthday. Distributed knowledge is a rather weak notion, but can be obtained of rather strong facts. Distributed knowledge is sometimes referred to as 'the wise man knows'.

2.3.2 Semantics

The semantics for our modal language $\mathcal{L}_{KC}$ can be obtained without adding additional features to our epistemic models $M = \langle S, \sim, V\rangle$ as defined in Definition 2.6. Recall that E_B has been discussed in Section 2.2.1 and can be defined within $\mathcal{L}_{KC}$. Let us also consider an operator D_B that models distributed knowledge in the group B. The language $\mathcal{L}_{KCD}$ extends $\mathcal{L}_K$ with this operator. The operators E_B, D_B and C_B are all necessity operators, and the accessibility relation that we need for each of them can be defined in terms of the relations $R_a(a \in A)$.

Definition 2.28 Let S be a set, and $R_b(b \in B)$ be a set of relations on it. Recall that a relation R_b is nothing but a *set* $\{(x, y) \mid R_bxy\}$.

- Let $R_{E_B} = \bigcup_{b\in B} R_b$.
- Let $R_{D_B} = \bigcap_{b\in B} R_b$.
- The *transitive closure* of a relation R is the smallest relation R^+ such that:
  1. $R \subseteq R^+$;
  2. for all x, y, and z, if $(R^+xy \& R^+yz)$ then R^+xz

  If we moreover demand that for all x, R^+xx, we obtain the *reflexive transitive closure* of R, which we denote with R^*. □

Note that R^*xy if y is *reachable* from x using only R-steps. More precisely, we have the following:

Remark

1. If R is reflexive, then $R^+ = R^*$.
2. R^+xy iff either $x = y$ and Rxy or else for some $n > 1$ there is a sequence $x_1, x_2, \ldots x_n$ such that $x_1 = x, x_n = y$ and for all $i < n$, Rx_ix_{i+1}. □

Definition 2.30 Let, given a set P of atoms and A of agents, $M = \langle S, \sim, V\rangle$ be an epistemic model, and $B \subseteq A$. The truth definition of $(M, s) \models \varphi$, with $\varphi \in \mathcal{L}_{KCD}$ is an extension of Definition 2.7 with the following clauses:

- $(M, s) \models E_B\varphi$ iff for all $t, R_{E_B}st$ implies $(M, t) \models \varphi$.
- $(M, s) \models D_B\varphi$ iff for all $t, R_{D_B}st$ implies $(M, t) \models \varphi$.
- $(M, s) \models C_B\varphi$ iff for all $t, R^*_{E_B}st$ implies $(M, t) \models \varphi$.

For $R^*_{E_B}$ we will also write R_{C_B}, or even R_B. Note that, if R_a is an equivalence relation, then $R_a = R^*_a$. □

Thus, everybody in B knows φ in s, if every agent $b \in B$ only considers states possible, from s, in which φ is true. Phrased negatively, $E_B\varphi$ does not hold as long as there is one agent who considers a state possible in which φ is false. And, φ is distributed knowledge in s if φ is true in every state that is considered a possible alternative to s by *every agent* in B. The idea being, that if one agent considers t a possibility, given s, but another does not, the latter could 'inform' the first that he need not consider t. Finally, φ is common knowledge of B in s, if φ is true in every state that is reachable from s, using any accessibility of any agent in B as a step. Again, put negatively, φ is *not* commonly known by B in s, if some agent in B considers it possible that some agent in B considers it possible that ... some agent in B considers it possible that φ is false.

Concerning the relation between the types of knowledge presented, one may observe that we have, if $a \in B$:

$$C_B\varphi \Rightarrow E_B\varphi \Rightarrow K_a\varphi \Rightarrow D_B\varphi \Rightarrow \varphi$$

which makes precise that common knowledge is a strong notion (which is easiest attained for weak φ, like tautologies), and distributed knowledge is weak (obtainable about strong statements, 'closest to the true ones').

Since the truth definition of $E_B\varphi$ should follow from its syntactic definition as given on page 12, and from now on we will rarely consider distributed knowledge, we define $\mathcal{S}5\mathcal{C}$ as the class of models in $\mathcal{S}5$ for which we have a truth definition for common knowledge:

Definition 2.31 The class of epistemic models $\mathcal{S}5\mathcal{C}$ will be $\mathcal{S}5$ with the truth definition for $C_B\varphi$, $(B \subseteq A)$, as given in Definition 2.30. □

Example 2.32 (Common knowledge in consecutive numbers) Consider the consecutive numbers example again (Example 2.4, the solution to Exercise 2.10, and, specifically, the model M from Figure A.1). Given that the actual numbers are $\langle 3, 2\rangle$, Bill obviously does not have a 4, although this is not clear for everyone: $M, \langle 3, 2\rangle \models \neg b_4 \wedge \neg E_{ab}\neg b_4$ (since $\neg K_b\neg b_4$ holds here). Also, although everybody knows that Anne does not have a 5, not everybody knows that everybody knows this: $M, \langle 3, 2\rangle \models E_{ab}\neg a_5 \wedge \neg E_{ab}E_{ab}\neg a_5$ (since $\langle 3, 2\rangle \sim_b \langle 3, 4\rangle$ and $\langle 3, 4\rangle \sim_a \langle 5, 4\rangle$ and $M, \langle 5, 4\rangle \models a_5$). We can continue and observe that $M, \langle 3, 2\rangle \models E_{ab}E_{ab}\neg b_6 \wedge \neg E_{ab}E_{ab}E_{ab}\neg b_6$.
The reader should be convinced now that, even though Anne sees a 2 on Bill's head, and Bill notices the 3 on Anne's head, it is not common knowledge between Anne and Bill that Anne's number is not 1235! (Bill considers it

possible he has a 4 in which case Ann considers it possible she has a 5 in which case Bill cannot rule out he has a 6 in which case ....) It *is* common knowledge between them though, given $\langle 3, 2\rangle$, that Bill's number is not 1235! (Since both agents only consider worlds possible in which Bill's number is even.) As a validity:

$$M \models (a_3 \wedge b_2) \rightarrow (\neg C_{ab}\neg a_{1235} \wedge C_{ab}\neg b_{1235}) \qquad \square$$

We give one final example concerning the semantics of common knowledge, in this chapter. It is a stronger version of the Byzantine generals, since its assumptions are weaker.

Example 2.33 (Possible delay) Two parties, S and R, know that their communication channel is trustworthy, but with one small catch: when a message Msg is sent at time t, it either arrives immediately, or at time $t+\epsilon$. This catch is common knowledge between S and R. Now S sends a message to R at time t_0. When will it be common knowledge between S and R that Msg has been delivered? Surprisingly, the answer is: "Never!"

Let us model this as follows. Write m for: Msg has been delivered, and let a state $\langle s_i, d_j\rangle$ be the state in which Msg has been sent at time t_i, and delivered at time t_j. So, it is commonly known that the difference between i and j is either 0 or ϵ. Assume that in the 'real' state Msg was sent at $i = 0$. For any time point $t_0 + k \cdot \epsilon$ ($k \in \mathbb{N}$), let M_k be the epistemic model representing the knowledge after k steps. Thus, $M_k, \langle s_0, d_0\rangle$ represents the epistemic state in which there was no delay in delivery of Msg and in which k time steps have passed, and $M_k, \langle s_0, d_1\rangle$ stands for the situation in which the message was sent at t_0, delivered after one delay ϵ, and after which k steps of epsilon have passed after t_0. It is important to realise that

$$M_k, \langle s_i, d_j\rangle \models m \text{ iff } j \leq k$$

If we assume that Msg is delivered immediately, and we investigate the situation at t_0, we get the model M_0 of Figure 2.5, with the real state $\langle s_0, d_0\rangle$. In this state, m is true, even R knows it (since the states that R considers possible, given $\langle s_0, d_0\rangle$, are $\langle s_0, d_0\rangle$ and $\langle s_{-1}, d_0\rangle$ (R knows Msg has been delivered, but he holds it for possible that this took ϵ delay, i.e., that it was sent at s_{-1}). In $\langle s_0, d_0\rangle$, agent S does *not* know m: he considers it possible that this takes a delay and that the real state is $\langle s_0, d_1\rangle$. So, we conclude that $M_0, \langle s_0, d_0\rangle \models \neg C_{SR}m$.

All the states in which m is true are those in the upper oval, in Figure 2.5. Note that we have not specified a beginning of time, in fact the 'upper half' can be finite, or even consist only of $\langle s_0, d_0\rangle$.

Let us now wait one ϵ, and obtain model M_1 of Figure 2.6. Of course, $\langle s_0, d_0\rangle$ is still a description of the actual state: we assumed that the message was sent at time 0, and immediately delivered. The states $\langle s_0, d_1\rangle$ and $\langle s_1, d_1\rangle$ now also verify m: since we are at time 1, and delivery in those states is at 1,

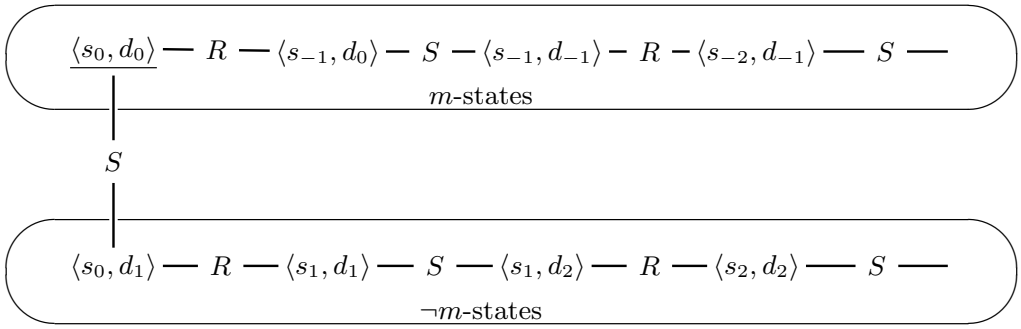

Figure 2.5. No delay, M_0.

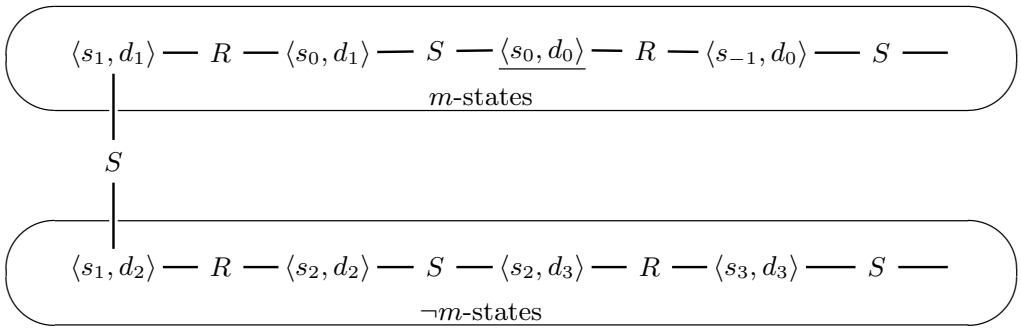

Figure 2.6. No delay, M_1.

these two states get 'promoted' to the 'upper oval' of the model. See Figure 2.6. We have $M_1, \langle s_0, d_0\rangle \models E_{SR}E_{SR}m$, but $M_1, \langle s_0, d_0\rangle \models \neg E_{SR}E_{SR}E_{SR}m$: given $\langle s_0, d_0\rangle$, sender S holds it possible that *Msg* was sent at 0 and delivered with delay ($\langle s_0, d_1\rangle$), a state in which R holds it for possible that the message arrived at time 1 without delay ($\langle s_1, d_1\rangle$), a situation in which S should be prepared to accept that the message is sent at time 1 *with* delay ($\langle s_1, d_2\rangle$). All in all, we conclude $M_1, \langle s_0, d_0\rangle \models \neg C_{SR}m$.

From all this, it should be clear that, no matter how many units k of ϵ we wait, we will get a model M_k in which the state $\langle s_0, d_0\rangle$ is 'shifted' $k + 2$ positions to the right, and the 'first' state down left the model M_k would be $\langle s_k, d_{k+1}\rangle$. In this state, at time k, delivery has not taken place, i.e., m is false, and, since $\langle s_k, d_{k+1}\rangle$ is $R_{C_{SR}}$-accessible from $\langle s_0, d_0\rangle$, we have $M_k, \langle s_0, d_0\rangle \models \neg C_{SR}m$ □

Exercise 2.34 What consequences would it have if the above guarantee would hold the way we send e-mail using the Internet? □

$C_B(\varphi \to \psi) \to (C_B\varphi \to C_B\psi)$	distribution of C_B over $\to$
$C_B\varphi \to (\varphi \wedge E_B C_B\varphi)$	mix
$C_B(\varphi \to E_B\varphi) \to (\varphi \to C_B\varphi)$	induction of common knowledge
From φ, infer $C_B\varphi$	necessitation of C_B

Table 2.3. Axioms of the epistemic system **S5C**.

2.3.3 Axiomatisation

The following definition establishes the axiomatisation of $\mathcal{S}5\mathcal{C}$.

Definition 2.35 Let A be a given set of agents, and let B be an arbitrary subset of it. Then the axiom system **S5C** consists of all the axioms and rules of **S5**, plus the axioms and derivation rule of Table 2.3. □

Axiom 'mix' implies that common knowledge is veridical. It also ensures all finite restrictions of the 'definition' of common knowledge given on page 31: one can show that this axiom and the definition of E_B ensure that, for every $k \in \mathbb{N}$, we have $\vdash C_B\varphi \to E_B^k\varphi$. The distribution axiom and the necessitation rule ensure that C_B is a normal modal operator. The induction axiom explains how one can derive that φ is common knowledge: by deriving φ itself together with common knowledge about $\varphi \to E\varphi$. We think that proving its soundness may help in understanding it.

Example 2.36 We show that the induction axiom is valid in the class of $\mathcal{S}5\mathcal{C}$-models. So, let M be an arbitrary $\mathcal{S}5\mathcal{C}$ model, then we have to prove that in any s, $(M,s) \models C_B(\varphi \to E_B\varphi) \to (\varphi \to C_B\varphi)$. To do so, assume that $(M,s) \models C_B(\varphi \to E_B\varphi)$ (1). This means that $(M,t) \models \varphi \to E_B\varphi$ for all t for which $R^*_{E_B}st$ (2). In order to prove the consequent of the induction axiom, suppose $(M,s) \models \varphi$ (3). Our task is now to show $(M,s) \models C_B\varphi$, which by Remark 2.29 is equivalent to saying: (4) for all n such that $R^n_{E_B}st$, we have $(M,t) \models \varphi$. And, indeed, the latter is done with induction over n. The basic case ($n = 0$) requires us to establish that for all t for which $R^0_{E_B}st$, i.e., for $t = s$, that $(M,t) \models \varphi$, which is immediate from (3). So, now suppose claim (4) is true for n. Take any t that is $n+1$ R_{E_B}-steps away from s, i.e. for which $R^{n+1}_{E_B}st$. Then there must be a state u for which $R^n_{E_B}su$ and $R^1_{E_B}ut$. The induction hypothesis guarantees that $(M,u) \models \varphi$, and from (1) we derive $(M,u) \models \varphi \to E_B\varphi$. This implies that $(M,u) \models E_B\varphi$, and hence, since $R^1_{E_B}ut$, also $(M,t) \models \varphi$, which completes the proof. □

The following exercise asks for some derivations in **S5C**. The first item establishes positive introspection of common knowledge, the second negative introspection. The third and fourth items demonstrate that an agent in B can never have any uncertainty about the common knowledge of B: $C_B\varphi$ holds if and only if some agent in B knows that it holds. Item 5 of Exercise 2.37 demonstrates that any depth of mutual knowledge of members in B about

φ follows from φ being common knowledge within B. Note that, by using contraposition, as soon as we have, for some chain of agents $a_1, a_2, \ldots, a_n \in B$ we can establish that $\hat{K}_{a_1}\hat{K}_{a_2}\cdots\hat{K}_{a_n}\neg\varphi$, then φ cannot be common knowledge in B. Finally the last item of Exercise 2.37 guarantees that common knowledge is preserved under subgroups.

Exercise 2.37 Let $a \in B \subseteq A$. Show that the following are derivable, in **S5C**:

1. $C_B\varphi \leftrightarrow C_BC_B\varphi$
2. $\neg C_B\varphi \leftrightarrow C_B\neg C_B\varphi$
3. $C_B\varphi \leftrightarrow K_aC_B\varphi$
4. $\neg C_B\varphi \leftrightarrow K_a\neg C_B\varphi$
5. $C_B\varphi \rightarrow K_{a_1}K_{a_2}\ldots K_{a_n}\varphi$, where every $a_i \in B (i \leq n)$
6. $C_B\varphi \rightarrow C_{B'}\varphi$ iff $B' \subseteq B$

Theorem 2.38 (Soundness and completeness) Compare. Theorem 7.19 For all $\varphi \in \mathcal{L}_{KC}$, we have $\vdash_{\mathbf{S5C}} \varphi$ iff $\mathcal{S}5\mathcal{C} \models \varphi$. □

2.4 Logics for Belief

Although there are many different logics for belief, it is generally well accepted that the main difference between knowledge and belief is that the latter does not comply with axiom T: whereas it does not make sense to say 'John knows today it is Tuesday, although it is Wednesday', it seems perfectly reasonable to remark 'John believes today is Tuesday, although in fact it is Wednesday'. So for beliefs (which we refer to using B_a) the property T, $B_a\varphi \rightarrow \varphi$, may fail.

Conversely, given a doxastic notion (i.e., belief) one may wonder what needs to be added to make it epistemic (i.e., knowledge). Obviously, for belief to become knowledge, it has to be true, but it is widely acknowledged that some additional criterion has to be met (suppose two supporters of opposite sport teams both believe that 'their' team will win the final; it would be not immediately clear that we should coin one of these attitudes 'knowledge'). In Plato's dialogue *Theaetetus*, Socrates discusses several theories of what knowledge is, one being that knowledge is true belief 'that has been given an account of'. Although in the end rejected by Socrates, philosophers have embraced for a long time the related claim that 'knowledge is true, justified belief'. A paper by Gettier, *Is Justified True Belief Knowledge?*, marked an end to this widely accepted definition of belief and sparked a lively and interesting discussion in the philosophical literature. Here, we will refrain from such deliberations.

From a technical point of view, to recover some kind of reasoning abilities of agents, it is often assumed that, although believed sentences need not be

all instantiations of propositional tautologies	
$K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$	distribution of K_a over $\to$
$\neg B_a\bot$	consistent beliefs
$B_a\varphi \to B_aB_a\varphi$	positive introspection
$\neg B_a\varphi \to B_a\neg B_a\varphi$	negative introspection
From φ and $\varphi \to \psi$ infer ψ	modus ponens
From φ infer $B_a\varphi$	necessitation of belief

Table 2.4. Axioms for the belief system **KD45**.

true, they should at least be *internally consistent*. In other words, for belief, T is replaced by the weaker axiom D: $\neg B_a\bot$. This is the same as adding an axiom D' : $B_a\varphi \to \neg B_a\neg\varphi$ (see Exercise 2.39): for any φ that the agent accepts to believe, he cannot also believe its opposite $\neg\varphi$ at the same time. For the other axioms, it is not so straightforward to pick a most obvious or even most popular choice, although belief systems with positive and negative introspection are quite common. Let us summarise the axioms of the belief system **KD45**, obtained in this way, in Table 2.4.

Exercise 2.39 Given that the axiom D' is $B_a\varphi \to \neg B_a\neg\varphi$, show that indeed axiom D and D' are equivalent with respect to **K**. □

The next two exercises show that, firstly, the system for belief is indeed weaker than that of knowledge, and secondly, how one can derive *Moore's principle* (2.4) in it. Moore's principle states that a rational agent (say, one whose beliefs are consistent), will not believe, at the same time, for any φ, that it is true while he does not believe it. This principle will play a main role in the next chapter's attempt to relate belief revision to dynamic epistemic logic, and it will also be a source of many paradoxical situations in both fields, as will become clear in the next chapters.

$$\neg B_a\bot \to \neg B_a(\varphi \wedge \neg B_a\varphi) \tag{2.4}$$

Exercise 2.40 Show that, indeed, T is stronger than the axiom D. That is, show that, $\vdash_{\mathbf{K}+T} D$, but not $\vdash_{\mathbf{K}+D} T$. For the latter, use that $\vdash_{\mathbf{K}+D} \varphi$ iff $KD \models \varphi$, where KD is the set of all serial Kripke models. □

Exercise 2.41 Show that Moore's principle (2.4) is derivable in $KD45$, by showing that even $\neg B_a(\varphi \wedge \neg B_a\varphi)$ has a derivation. □

We mention the following result.

Theorem 2.42 Axiom system **KD45** is sound and complete with respect to the semantic class $\mathcal{KD}45$, i.e., for every formula φ, we have $\vdash_{\mathbf{KD45}} \varphi$ iff $\mathcal{KD}45 \models \varphi$. □

Of course, it may be interesting to study frameworks in which *both* knowledge and belief occur. Although $K_a\varphi \to B_a\varphi$ is an accepted principle relating

the two (fuelled by the claim 'knowledge is justified, true belief'), it is not immediately clear which other interaction properties to accept, like for instance $K_a\varphi \to B_aK_a\varphi$. Also, one can, as with knowledge, define group notions of belief, where in the case of belief common belief is usually interpreted with respect to the transitive closure (note: not the reflexive transitive closure) of the union of the individual accessibility relations.

2.5 Notes

Hintikka, most notably through [99], is broadly acknowledged as the father of modern epistemic logic, although Hintikka himself thinks that von Wright deserves this credit.[2] Modern epistemic logic started to flourish after modal logic (with its roots in Aristotle) was formalised and given a possible world semantics. It is hard to track down the exact origins of this semantics, but it is widely known as Kripke semantics, after Kripke, who devoted a number of early papers to the semantics of modal logic (see [120]). A contemporary and thorough standard work in modal logic is the monograph [29] by Blackburn, de Rijke and Venema, to which we refer the reader for a deeper analysis and for further references on this subject.

From the late 1970s, epistemic logic in the sense it is treated in this chapter became subject of study or applied in the areas of artificial intelligence (witnessed by Moore's early work [152] on reasoning about actions and knowledge), philosophy (see Hintikka's [100]), and game theory. Regarding the latter, Aumann is one of the most prominent to mention. His [5] gives one of the first formalisations of common knowledge, a notion that was already informally discussed in 1969 by Lewis in [128]. And in Aumann's survey paper [6] on interactive epistemology the reader will immediately recognise the system $\mathcal{S}5$. Together with Brandenburger, Aumann argued in [7] that knowledge is crucial for game theoretic solutions. For a contemporary and modal logical treatment of the latter, see also de Bruin's thesis [33].

In the 1980s, computer scientists became interested in epistemic logic. In fact, the field matured a lot by a large stream of publications around Fagin, Halpern, Moses, and Vardi. Their important textbook *Reasoning about Knowledge* [62] which appeared in 1995, is in fact a survey of many papers co-authored by (subsets of) them over a period of more than ten years. We refer to [62] for more references. Their emphasis on *interpreted systems* as an underlying model for their framework makes the **S5** axioms easy to digest, and this semantics also facilitates reasoning about knowledge during *computation*

[2] Hintikka referred to von Wright as the founder of modern epistemic logic in his invited talk at the PhiLog conference *Dimensions in Epistemic Logic*, Roskilde, Denmark, May 2002. In the volume [96] dedicated to this event Hintikka [101] refers to von Wright's [195] as the thrust of epistemic logic (which was practiced already in the Middle Ages, see [30]) 'to the awareness of contemporary philosophers'.

runs in a natural way. Their work also generated lots of (complexity) results on knowledge and time, we also mention the work of van der Meyden (e.g., [143]) in this respect. The textbook [148] on epistemic logic by Meyer and van der Hoek also appeared in 1995. There, the emphasis is more on 'classical Kripke-world semantics', and the book discusses various notions of belief, and non-classical reasoning. All the theorems mentioned and the proofs omitted in this chapter can be found in [62, 148].

In the 1990s, the paradigm of *agents* re-enforced the computer science community's interest in modal logical approaches to notions like knowledge and belief. We here only mention the influential BDI ('Belief, Desires, and Intentions') framework, as developed by Rao and Georgeff [172]. For a more extensive survey on modal logics for rational agents, see van der Hoek and Wooldridge's [105] and the references therein. Recently, we see a growing interest in epistemic knowledge by researchers in computer science, agent theory and games. Such a multi-disciplinary stance was much influenced by the TARK [188] and LOFT [136] conferences: we refer to their proceedings for further references.

Regarding Example 2.3, the proof that knowledge of depth four is sufficient and necessary to comply with the protocol can be found in Halpern and Zuck's [91], which gives a thorough knowledge-based analyses of the so-called sequence-transmission problem. Meyer and van der Hoek [148] give a detailed description of how to transfer the solution given here into the 'alternating bit protocol', in which all explicit references to knowledge have been removed. The aim of this example is to show how a knowledge-based specification can help to find a solution: the alternating bit protocol itself was known before epistemic logicians thought about this kind of problems (see Bartlett, Scantlebury, and Wilkinson's analyses [13] of 1969). For a simulation of the protocol, the reader be referred to `http://www.ai.rug.nl/mas/protocol/`.

Although [148] gives a rather ad-hoc procedure to eliminate the epistemic operators from the specification, there is a stream of research in *knowledge-based programs* (first defined by Kurki-Suonio [122]) that tries to systematically determine which kind of program can 'count as' a knowledge-based program: see the work by Halpern [84], Vardi [190], or their joint work with Fagin and Moses [63]. Example 2.3 in fact analyses a *Knowledge-based protocol*, a term that was introduced in 1989 by Halpern and Fagin [86], and which is still popular in analysing for instance protocols for the Internet (see Stulp and Verbrugge's [187] for an example).

Our running example of consecutive numbers is one of the many and popular puzzles trying to explain common knowledge, like the muddy children puzzle (see Chapter 4). The original source of the consecutive number example is not known to us, but the earliest reference we found to a variant of this puzzle is in Littlewood's [135] from 1953. A recent analysis of the problem using *Interactive Discovery Systems* can be found in Parikh's [162], with references to Littlewood and to van Emde Boas, Groenendijk, and Stokhof's [60]. Littlewood's version presented in [31], which is an extended rewriting of [135], reads as follows:

> The following will probably not stand up to close analysis, but given a little goodwill is entertaining.
> There is an indefinite supply of cards marked 1 and 2 on opposite sides, and of card marked 2 and 3, 3 and 4, and so on. A card is drawn at random by a referee and held between the players A, B so that each sees one side only. Either player may veto the round, but if it is played the player seeing the higher number wins. The point now is that every round is vetoed. If A sees a 1 the other side is 2 and he must veto. If he sees a 2 the other side is 1 or 3: if 1 then B must veto; if he does not veto then A must. And so on by induction.

The notion of bisimulation is a focal point in expressivity results of modal logic, and was independently discovered in areas as diverse as computer science, philosophical logic, and set theory. In computer science it plays a fundamental role in the analysis of when two automata 'behave the same'. In this context, the notion of bisimulation was developed through a number of co-inspired papers by Park [165] and Milner [150]. Van Benthem introduced [18] p-morphisms (essentially bisimulations) in his correspondence theory for modal logic, and finally, Forti and Honsell developed their notion of bisimulation for their work [65] on non-well founded sets. A nice and short overview of the origins of bisimulation is given in Sangiorgi's [180].

The issue of logical omniscience was already mentioned in [99] and quite extensively discussed in the monographs [62] and [148], with several solutions mentioned. Further references are found there.

Completeness proofs for the logics presented here can be found in many textbooks. A standard reference for techniques in modal logic is Blackburn, de Rijke, and Venema's [29]. Strong completeness and the role of applying necessitation to the premises is discussed in papers by van der Hoek and Meyer [106] and by Parikh [163]

The problem of the Byzantine generals was formalised already in 1980 by Pease, Shostak, and Lamport [166]. A survey of this problem is provided by Fischer's [64]. The formalisation of the 'possible delay' problem, (Example 2.33), is taken from [62], although their treatment is in the setting of interpreted systems. Our semantic analysis is inspired by van der Hoek and Verbrugge [104].

The famous paper 'Is Justified True Belief Knowledge?' by Gettier [78] is a good example of a clear challenge of one definition of belief, and has many successors in the philosophical literature. Kraus and Lehmann [119] were among the first who looked at logics that combined knowledge and belief defined in a possible worlds framework. Van der Hoek [102] gives a semantic analysis of 'how many' interaction axioms (like the mentioned $K_a\varphi \rightarrow B_aK_a\varphi$) one can assume without the two notions of knowledge and belief collapsing to one and the same thing.