Chapter 5

Knowledge and Information Flow

Overview The first part of this course has shown how logical systems describe the world using objects, predicates, quantifiers and propositional combinations. This information about the world is typically conveyed when we use language, and such information then leads to knowledge among language users. This chapter deals with the logic of knowledge as based on information, including changes in knowledge which result from observations of facts, or communication between agents knowing different things. This area is called *epistemic logic*, and its main difference with the earlier systems of Chapters 2, 3 and 4 is that we can also express facts about knowledge of one or more agents in the logical language itself. This 'social' perspective occurs in many settings: knowing what others do or do not know determines our actions. Another central theme of this chapter is "change": successive information processing steps change what agents know, and this, too, is essential to understanding the logic of language use and other cognitive tasks.

5.1 Logic and Information Flow

From truth to informational actions by agents In this course, we have explained a valid argument such as

$$\text{from } p \to q \text{ and } \neg q \text{ to } \neg p$$

as:

> whenever $p \to q$ is true and $\neg q$ is true, $\neg p$ is also true.

But, true for whom? If we think about how a logical system is used, there is usually a knowing subject performing the inference, giving it a character more like this:

> If I know $p \to q$ and I know $\neg q$, then I also know $\neg p$.

And there need not be just me. Suppose I know that $p \to q$, while you do not know this, but you do know that $\neg q$. Then we should be able to pool our knowledge to reach the conclusion $\neg p$. What informational actions are involved here?

Recall what we explained in the introductory chapter of this book. Information can come from different events. The three main sources that have long been recognized for this are

observation, inference and *communication*.

Agents may come to know directly that propositions are true by perception, they can infer propositions from things they already know, and they can also learn things from others.

Example 5.1 Here is a story which ancient Chinese logicians used already some 2500 years ago to make this point:

> Someone is standing next to a room and sees a white object outside. Now another person tells her that there is an object inside the room of the same colour as the one outside. After all this, the first person knows that there is a white object inside the room. This is based on three actions: an observation, then an act of communication, and finally an inference putting things together.

Being explicit about what agents know The logical systems that you have learnt in this course can express the factual content of the information that we have just discussed. They can even model information flow through the steps of inference or update in propositional logic. But if we want to bring more of this information flow into a logical system, we need to talk about agents. This is even more pressing since we often reason explicitly about other agents' knowledge. We ask questions to people when we think they may know the answer, and we are also interested in what other people do not know, like when a teacher is trying to tell the students something new.

All this may sound simple and familiar, but it also poses interesting challenges. I tell you: "You *don't know* it, but these days I live in Seville." Having now told you this, you *know* that I live in Seville! So, what I told you has become false precisely because I said it. Is this not a paradox? One of the things you will see in this chapter is why this is not paradoxical, but rather a typical point in the logic of communication.

This *epistemic logic* of this chapter starts from propositional logic, but now enriched with logical operators $\Box_i \varphi$ (also written als $K_i \varphi$) standing for the natural language expression "agent i knows that φ". Once we have this system in place, we will also use it for a description of information flow by observation and communication.

The subjects whose knowledge we describe are called *agents*. These agents can be human beings, but also measurement devices, or information processes running on a computer, each with their own informational view of the total situation. A system modeling the interaction of different agents is often called a *multi-agent system* . The agents of such

a system can be a group of people having a conversation, but also the many computers making up the Internet, or in more physical terms, the players of a soccer team with their varying fields of vision and varying abilities to change their positions.

Exercise 5.2 Imagine players in a soccer match. What channels of information are available to them? What factors restrict the information flow? (If you don't like soccer, then you are invited to replace this example with your own favourite game.)

5.2 Information versus Uncertainty

Before we give our logical system, let us first ask a preliminary question:

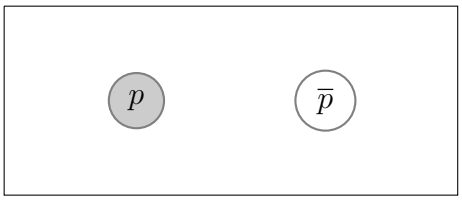
What is information?

Actually, this is not easy to say, and many disciplines in the university have things to contribute. But here is a basic idea that occurs widely: we approach things in the converse order, and model not information but *uncertainty*. And the idea for that is simple. To a first approximation, this idea can be stated with the semantic valuations for propositional logic that you have learnt. There is one actual world or actual situation, but we may not yet know what it is, and hence we must consider a larger range of options:

Uncertainty is the set of current options for the actual world.

Information models: sets of possible situations What these options are depends on the concrete scenario that we have in mind. Let us look at some examples.

Example 5.3 (Two options for one agent.) You are about to kick a ball. You may either score a goal, or not. The two options 'score', 'not-score' are the two relevant situations, which you can represent as the two valuations for an atomic statement p = 'I score a goal', one with $V(p) = 1$ and one with $V(p) = 0$. The same pattern can of course happen in many other scenarios: 'Pass' or 'Fail' for an exam that you have taken, 'Head' or 'Tails' for the outcomes of the next throw of a coin, 'Left' or 'Right' for the correct way to the Rijksmuseum, and so on. It is often easiest to think of this in the form of a *picture*. In what follows, the circles stand for the possibilities (for which we also use the term *worlds*), the proposition letters indicate which atomic facts are true where, and the shaded circle indicates the actual situation.

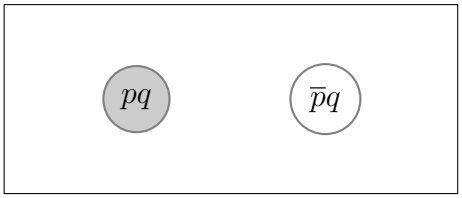


In this picture, p is true, but there is uncertainty about this: the actual situation cannot be distinguished from a situation where p is false (indicated by $\overline{p}$). Later on in this chapter, we will be more precise about the definition of such pictures, and how they can be viewed as information models.

So far, we have left something out that is often important to make explicit. We assume that there always is an *actual situation*, the reality we live in. That means that one of the possibilities in the model is the 'actual' or 'real' one. Of course, the agent herself need not know which of the various possibilities this is: if she did, she would know more than what we have pictured! One often indicates this actual world in some special way. Think of it as a marking that is invisible to the agents inside the model. It may be put there by an omniscient outside observer, or by you as the designer of the scenario. In our pictures, the actual world is usually marked by shading it grey. In the above picture, therefore, the actual world is the one to the left, and the truth of the matter is that p holds.

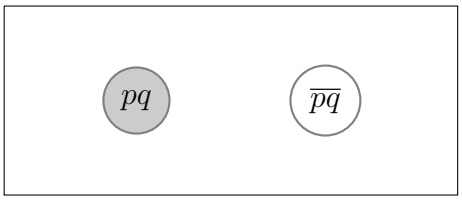
It seems that there was no information at all in the preceding model. But things soon get more interesting.

Example 5.4 (Ignorance and knowledge.) Suppose that your current range of options is the two valuations $\{V, V'\}$, with $V(p) = V(q) = 1$ and $V'(p) = 0$, $V'(q) = 1$.



Intuitively, you still know nothing about p, but you do *know that q is the case*, since it is true in every possibility that you consider.

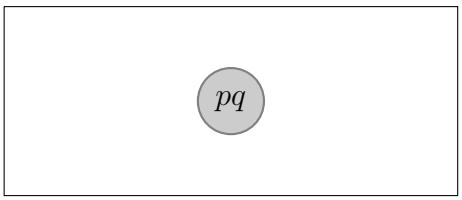
Or things could even be a bit more delicate. This time, your current range of options is again two valuations $\{V, V'\}$, but now with $V(p) = V(q) = 1$ and $V'(p) = V'(q) = 0$.



Now you do not know whether p is true, or whether q is true, but you do know that one is true if and only if the other is. In particular, if you can find out the truth value of p, say by making an observation or asking some trusted person, you will automatically know the truth value for q. This is indeed how information works: Suppose you know that Mary

and John are a dancing couple, and that either both of them show up in the ballroom or neither. If someone tells you that they saw Mary, you conclude that John was there as well.

As a very special case, a model with just one option represents a situation of complete information about what things are really like. In the following picture, the agent knows what the actual world is, and in particular, that both p and q are true there:



One set of possibilities, as we have considered so far, just describes what one particular agent knows and does not know. Now let us look at a typical situation where more agents are involved. Suppose that

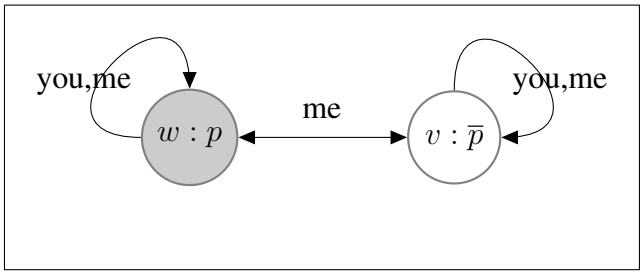
> I do not know if p is the case, but I know that you know *whether* p is the case (i.e., if p is true, you know that, and if p is false, you know that, too).

This would be a good reason, for instance, for me to ask you a *question*, namely, "Is p the case?". How can we model this social scenario?

Example 5.5 (Two options and two agents.) There are two possibilities for whether p is the case or not; let us call them p, $\overline{p}$ for short. But this time, there is a distinction between how I see them and how you see them. For me, both options are possible, so they belong to the same set $\{p, \overline{p}\}$. To indicate that I cannot distinguish between these cases, we can draw a connecting arrow marked 'me'. But for you, the two possibilities do not belong to the same set, since your knowledge actually allows you to distinguish them. That is, you distinguish the two sets $\{p\}$ and $\{\overline{p}\}$. For you, there is no connecting arrow between the two possibilities.

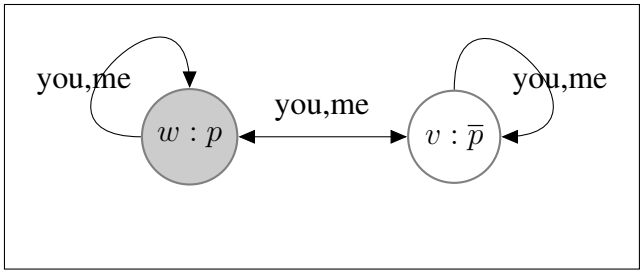
Now, a picture really helps to visualize what is going on. We draw the two relevant options as points w, v with the truth or falsity of p indicated.

But this time there is a new feature. The *line with the two arrowheads* in the picture indicates which situations agent 'me' cannot distinguish. More precisely, I see w connected to v by a line marked with the 'label' *me* because both are options for me. But you have no line there, since you are informed about the distinction between p and $\overline{p}$.



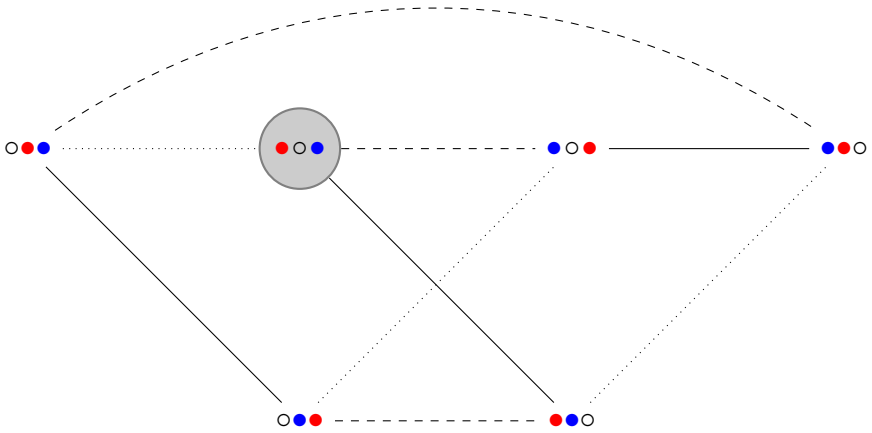
What about the loop arrows drawn for you and me? These indicate that we cannot distinguish an option from itself. That sounds obvious, but we will soon see what this means for the logic of knowledge. (In the rest of this chapter, to avoid cluttering pictures, we will often drop such loop arrows, taking them as tacitly understood.)

If we were to add a labeled arrow between the two worlds for the agent you to this picture, the model would show that neither agent knows if p is the case:



Beyond these simple examples, more interesting scenarios arise in simple stories with *cards*. Parlour games are a concrete "information lab" with agents knowing different things.

Example 5.6 Consider the start of a very simple card game. Three cards red, white, blue are given to three players: 1, 2, 3, one each. Each player sees her own card, but not that of the others. The real distribution over the players 1, 2, 3 (the "deal") is red, white, blue (●○●, or rwb). Here is a picture of the information model:



To avoid cluttering the picture, self loops and arrow points are omitted (but assumed to be there). The solid links are the links for player 1, the dashed links are the links for player 2, and the dotted links are the links for player 3. The 6 worlds are the 6 relevant options, which are the 6 possible deals of the cards (3 cards over 3 people), with the appropriate uncertainty links between deals, marked for players. For instance, the solid line between ●○● (rwb) and ●●○ (rbw) indicates that player 1 cannot distinguish these two situations, whereas 2 and 3 can: if 1 has the red card, she is uncertain about whether 2 and 3 have white and blue or blue and white respectively; however, there is no such uncertainty for 2 and 3, because 2 and 3 both know which card they have themselves.

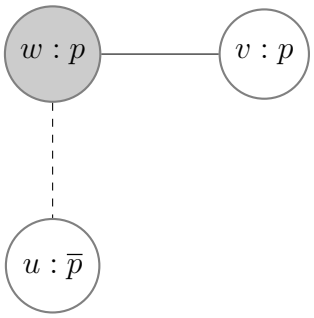
In words, the diagram says things like the following. In the actual world, agent 1 knows that she has the red card, since this is the case in the only two options that she considers possible. About the other agents, the only thing agent 1 knows is that they must have either blue or white, but she does not know in which order. As for the status of the shaded world ●○● (rwb): though the players are actually in ●○● (as an outside observer could see), none of 1, 2, 3 knows this. So, if they are to find out more, further information will have to come in. That is of course precisely the point of card games, and we will continue with this example later on.

You may already have recognized the structures that we are drawing here. They are *undirected graphs* in the sense of Chapter 4, consisting of points (the possibilities) and uncertainty lines for agents. So far, these lines have no direction: two points being indistinguishable is a symmetric relation. But in modeling other forms of information, it also makes sense to have directed graphs with arrows, as we will see later.

These examples were simple, but they should give the main idea. For now we conclude with a slightly more tricky example. We have suggested that possibilities are like propositional valuations. But this is not always completely true.

Example 5.7 (Knowledge about others.) Consider the following information model with two kinds of links for different agents. The solid links are for agent 1 and the dashed links

for agent 2.



Here the two worlds w, v have the same valuation with p true. And yet they are not identical. Why? Look at the actual world w. Agent 1 sees only the two worlds w, v there (indicated by the solid line), and hence she knows that p. But she is uncertain about what agent 2 knows. In world w, agent 2 does not know that p (he also considers the $\neg p$-world u possible, as indicated by the dashed line), but in world v, agent 2, too, knows that p (he has no uncertainty lines connecting world v to other worlds). Since w is the actual world, the truth of the matter is that 1 knows that p, and 2 does not, but 1 is not sure about what 2 knows. In particular, 1 considers it possible that 2 knows that p, but 1 also considers it possible that 2 does not know that p. How could such a situation come about? Well, for instance, 1 knows that p because she heard the teacher say that p, but she was not sure if the other student 2 was paying attention.

Many situations in real life are like this. Maybe you already know that this stove is hot, maybe you do not, but I know that it is. For safety's sake, I now tell you that it is hot, to make sure you do not hurt yourself. We will soon see how to reason about such subtle differences in information in a precise manner.

Exercise 5.8 Find a concrete practical situation where what matters to my actions is three repetitions of knowledge about other agents: what I know about what you know about my knowledge.

Information and knowledge in natural language We have given pictures for information of one or more agents. But of course we also use natural language to talk about information, both about the facts and about others. This is just the point of expressions like "John knows that Mary has taken his car", or "John does not know that Mary knows his pin code". We can introduce a convenient shorthand notation for this as follows:

$\Box_i \varphi$ for 'agent i knows that φ'.

An alternative for this is to use $K_i \varphi$ for $\Box_i \varphi$. The letter K stands for the k in the word "knowledge". In this book we will use $\Box_i$ rather than K_i. The operator $\Box_i$ comes with a dual $\Diamond_i$ (see below). (If there is just a single agent we use $\Box \varphi$ for "the agent knows that φ".)

This new operator $\Box_i$ can be freely combined with the logical languages that you have already learnt to form many types of expression involving information. For example, John (j) knows *whether* Mary has taken his car (p) if he knows which is which, so this sentence corresponds to the formula

$$\Box_j p \vee \Box_j \neg p.$$

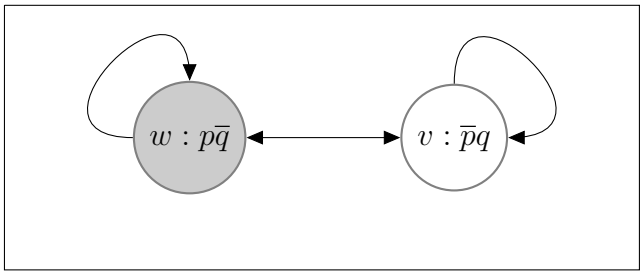
And reading formulas back into language, the formula

$$\Box_m(\neg\Box_j p \wedge \neg\Box_j \neg p)$$

says that Mary knows that John does not know if she has taken the car.

Example 5.9 (Knowledge and Negation) Note the difference between $\Box_j \neg p$ and $\neg\Box_j p$. The first expresses (under the above reading of p) that (Mary did not take the car and) John knows that Mary did not take the car, the second that (even if Mary took the car) John does not know that she took the car.

Example 5.10 (Knowledge and Disjunction) Knowing a disjunction is different from either knowing one disjunct or knowing the other disjunct. $\Box(p \vee q)$ is true in the actual world of the following model:



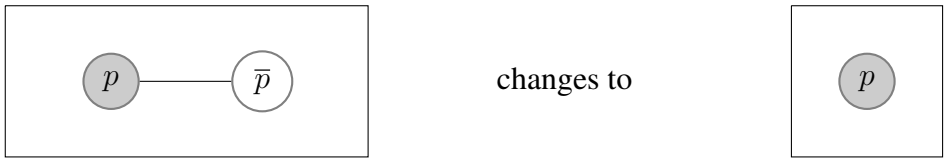
But $\Box p \vee \Box q$ is *not* true in the actual world of this model, for $\Box p$ and $\Box q$ are both false in the actual world.

Natural language has a whole range of expressions for information. We know that certain things are true, but we also know objects like telephone numbers, or methods for doing something. And other than knowing facts, we can also just "believe" them (a weaker informational notion), or we can merely consider them possible, doubt them, and so on. And these expressions are associated with actions that have names, too. You come to know facts by *learning* about them, an action that changes your information state, or as a teacher, you can *tell* other people what you know, another dynamic action. We are all experts in this informational repertoire, and it is hard to imagine life without it.

5.3 Modeling Information Change

What has been in the background of our examples all the time is that our information is not static, but that it typically keeps changing. How does this change happen? The basic idea is very simple: more information means reduction of uncertainty. And this reduction is achieved by means of shrinking the range of options, as we will see in the following pictures.

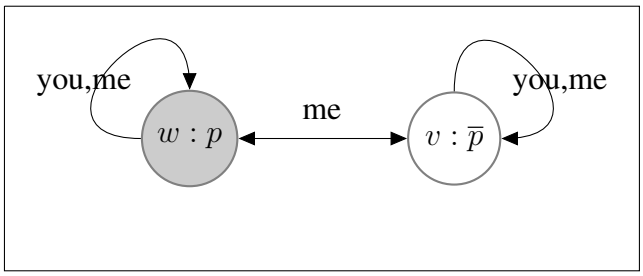
Example 5.11 (Finding out) I am uncertain whether it is raining (p) or not. I go to the window, look outside, and see that it in fact is raining. Here is what happens then to the earlier model:



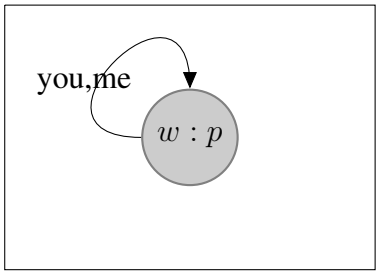
In the new smaller model, I know that p. The step taking me from the first model to the second is often called an *update*.

The same mechanism applies to acts of communication.

Example 5.12 (Answering a question.) Suppose that I do not know if p, but you do. I ask you the question "Is p the case?", and you answer truthfully. What happens when your answer? This time, the earlier model



changes to



where we both know that p. Actually, intuitively, we also know that we both know that p, since we are both in the very same situation. For more about this claim of information about other people's information, see below.

Our final example is the earlier simple card game, whose initial model was drawn above. Now the information flow already gets a bit less straightforward.

Example 5.13 (Finding out about Cards; compare Example 5.6) Suppose that the following two conversational moves take place between players:

(1) 2 asks 1: "Do you have the blue card?".

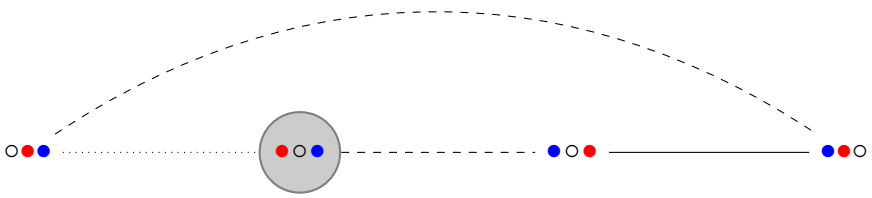
(2) 1 answers truthfully: "No".

Who knows what then? Here is the effect in words:

> Assuming the question is sincere, 2 indicates (just by asking it) that she does not know the answer, and so she cannot have the blue card. This tells 1 at once what the deal was. But 3 does not learn anything from 2's question, since he already knew that 2 does not have blue. When 1 says she does not have blue, this now tells 2 the deal. However, 3 still does not know the deal even then.

We now give the intuitive updates in the diagram, making the reasoning geometrically transparent. Here is a concrete update video of the successive information states:

After 2's question, the two situations where 2 has the blue card are removed, for the question reveals to everybody that 2 does *not* have the blue card.



After 1's answer "no", the situations where 1 has the blue card get removed, and we get:

We see at once in the final diagram that players 1, 2 know the initial deal now, as they have no uncertainty lines left. But player 3 still does not know, but she does know that 1, 2 know, and 1 and 2 know that 3 knows that 1 and 2 know, and 3 knows that 1 and 2 know that 3 knows that 1 and 2 know, and so on. In fact, that 3 knows that 1 and 2 know is *common knowledge* between the three agents: see Example 5.18 below for more information.

Note how the flow of information can be of different sorts: directly about the facts, or about what others know. Agent 3 did not find out what the deal is, but he did learn something that can also be of great interest, namely that the other players know the deal.

Exercise 5.14 (Misleading Questions.) In the preceding version of the scenario, we assumed that the question was "honest", and not misleading. That is why it gave everyone the reliable information that 2 did not have the blue card. Give examples of scenarios where questions do not indicate that the questioner does not know the answer. What if we drop this assumption in our game? What information flows then in the above scenario? What is the final diagram, and what do players know there?

Similar analyses of information flow exist for a wide variety of puzzles and games. Not all update steps are always of the above simple kind, however. We will briefly discuss more "private" forms of information flow in Outlook Section 5.10 to this chapter.

By now, we have raised a lot of issues about information and update in an informal manner. It is time to present a logical system that can deal with all this.

5.4 The Language of Epistemic Logic

We will now present the system of epistemic logic. What is important for you to see is that, even though the motivation given in this chapter may have sounded very different from that for propositional and predicate logic, the system that follows actually has a very similar technical structure. You will encounter all the topics that you have seen before: formal language, semantics, validity, proof, and update. Indeed, we will even be a bit shorter than in earlier chapters, since you can apply the techniques that you already know from there.

We start with defining the above notation more precisely.

Definition 5.15 (Basic epistemic language EL) Fix a set P of proposition letters, and a set I of agents. The basic epistemic language EL extends the language of propositional logic with modal operators $\Box_i\varphi$ ('i knows that φ'), for each agent $i \in I$. The inductive syntax rule is as follows, where p stands for any choice of proposition letters and i stands for any agent.

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid (\varphi \vee \varphi) \mid \Box_i\varphi \mid \Diamond_i\varphi.$$

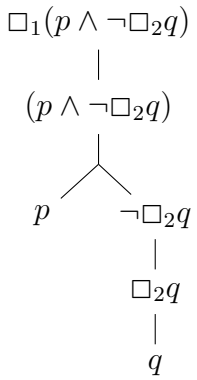
Notice that this definition covers all operations from propositional logic. We do not 'officially' put $\rightarrow$ and $\leftrightarrow$ into the language, but these can easily be defined. We know from propositional logic that $\varphi \rightarrow \psi$ can be expressed equivalently as $\neg(\varphi \wedge \neg\psi)$, and that $\varphi \leftrightarrow \psi$ can be defined as $(\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$. Henceforth we will use these abbreviations without further comment.

$\Box_i$ expresses knowledge for agent i, and $\Diamond_i$ expresses 'not knowing that not'. As the truth definition (see 5.24 below) will make clear, $\Diamond_i\varphi$ is equivalent to $\neg\Box_i\neg\varphi$. This says intuitively that 'agent i considers φ possible', or that 'according to i, φ cannot be ruled out'. The operator $\Diamond_i$ is called the *dual* of the operator $\Box_i$, just like the existential quantifier $\exists$ is the dual of the universal quantifier $\forall$.

Example 5.16 The following expression is a formula:

$$\Box_1(p \wedge \neg\Box_2 q).$$

Here is a construction tree in the same style as the ones you have seen in Chapters 2 and 4, showing the construction of the formula by following the above syntax rules step by step:



The following expressions are not formulas: $p\Box q$, $p\Box_1 q$, $\Box pq$.

Exercise 5.17 How many different correct formulas can you make from the following sequence of symbols by putting in brackets at appropriate places?

$$\neg\Box_i p \rightarrow q.$$

Also write the corresponding analysis trees for these formulas. Can you also explain what these different formulas say?

We have already shown informally how epistemic formulas match concrete natural language expressions. Here is a more elaborate example showing how this works:

Example 5.18 (Questions and Answers) I approach you in Amsterdam, and ask "Is this building the Rijksmuseum?". As a helpful Dutch citizen, you answer truly: "Yes". This is the sort of thing we all do all the time. But subtle information flows. By asking the question, I convey to you that I do not know the answer, and also, that I think it is possible that you do know. This information flows before you have said anything at all. After that, by answering, you do not just convey the topographical fact to me that this building is the Rijksmuseum. You also make me know that you know, and that you know that I know you know, etcetera. Even such a simple episode of a question followed by an answer mixes factual information with social information about the information of others. The latter type of information is not a mere "side-effect" of communication: it can steer further concrete actions. If you know that I know this building is the Rijksmuseum, and you see me running into that building waving a spray can, you may want to take some fast action.

Here is how our language can formulate some of the relevant assertions:

(i) $\neg\Box_Q\varphi \wedge \neg\Box_Q\neg\varphi$ 'questioner Q does not know whether φ',

(ii) $\Diamond_Q(\Box_A\varphi \vee \Box_A\neg\varphi)$ 'Q thinks it is possible that A knows the answer'.

After the whole two-step communication episode, φ is known to both agents:

(iii) $\Box_A\varphi \wedge \Box_Q\varphi$,

while they also know this about each other:

(iv) $\Box_Q\Box_A\varphi \wedge \Box_A\Box_Q\varphi$, $\Box_A\Box_Q\Box_A\varphi \wedge \Box_Q\Box_A\Box_Q\varphi$, etcetera.

This mutual knowledge (knowledge about knowledge of the other) to every finite depth of iteration is a property of the group $\{Q, A\}$ of the two agents together. It is called *common knowledge*.

We will return to group knowledge that arises from communication in the Outlook Section 5.11 at the end of this chapter.

Exercise 5.19 Give a concrete setting where questions have neither of the two forms of information that we mentioned in the preceding example.

Just as you have learnt with predicate logic, moving back and forth between natural language and formulas is something that you can practice.

Exercise 5.20 Write formulas that match the following sentences:

(1) John knows that it is not raining.

(2) John knows whether Mary knows that is is raining.

(3) John knows whether Mary knows if is is raining.

Exercise 5.21 Read the following formulas as sentences in natural language (pick a suitable key):

(1) $\Box_1(p \to \neg\Box_2 q)$,

(2) $\Box_1\Box_2 p \to \Box_2\Box_1 p$.

5.5 Models and Semantics for Epistemic Logic

Information models Now we state the formal version of the intuitive idea of information as range in our informal explanations. The following structures consist of a range W of 'worlds standing for all the options that we consider, while the earlier idea of lines for uncertainty is now stated formally as the presence of so-called 'accessibility relations' marked for the agents. A relation $w \to_i v$ (an arrow marked with agent i pointing from w to v) holds between two worlds if,

from the viewpoint of world w, agent i considers v a possible alternative.

Moreover, for each world, we mark which atomic facts are true there using a 'valuation'.

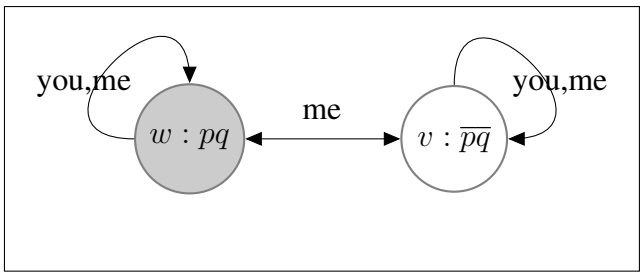
Definition 5.22 (Information Models) Models M for the epistemic language are triples

$$(W, \{\to_i |\ i \in I\}, V),$$

where W is a set of worlds, the $\to_i$ are binary accessibility relations between worlds, and V is a valuation assigning truth values to proposition letters at worlds. In what follows, we mostly use pointed models (M, s) where s is the actual world representing the true state of affairs.

Epistemic models may be viewed as collective information states for a group of agents. For illustrations, just think of all the examples we have seen already.

Example 5.23 (A Diagram Viewed as a Model) Here is a diagram that might describe the start of a question and answer episode (compare Example 5.4).



This is the following model $M = (W, \{\rightarrow_{\text{me}}, \rightarrow_{you}\}, V)$. The set W of worlds equals $\{w, v\}$, the valuation V is given by $V(w)(p) = 1$, $V(w)(q) = 1$, $V(v)(p) = 0$, $V(v)(q) = 0$, the relation $\rightarrow_{\text{me}}$ equals $\{(w, w), (w, v), (v, w), (v, v)\}$, and the relation $\rightarrow_{\text{you}}$ equals $\{(w, w), (v, v)\}$. As a pointed model, it has the shape (M, w), indicating that w is the actual world.

One way of thinking about what is happening here is that a model is not just one valuation for atomic facts, as in propositional logic, but a family of these. (This is not quite true, as we have seen in an earlier example, where different worlds could still have the same valuation for atomic facts. But often, the analogy is close enough.) Knowledge statements then refer, not just to one valuation, but to a whole range of them, as many as the agents' information requires. The case of just one world with its valuation then corresponds, as you have seen earlier, to perfect information by everybody concerning the actual world.

In general, we allow every sort of binary accessibility relations on epistemic models. Thus, any directed graph with a family of relations (indexed for the agents) could be used. We will give such a general picture in a moment. However, in practice, we often work with very special relations, of a sort already mentioned in Chapter 4 on predicate logic. These are so-called *equivalence relations*, satisfying the following three conditions:

reflexivity For all w, Rww.

symmetry For all w, v: if Rwv, then Rvw.

transitivity For all w, v, u, if Rwv and Rvu, then Rwu.

One can think of such relations as partitioning the total set of worlds into a number of disjoint maximal 'zones' of worlds connected by the relation. For instance, in the above example, the partition for Me has just one zone: the whole set $\{w, v\}$, while that for You has two zones: $\{w\}$ and $\{v\}$. This is easy to visualize, and we will soon discuss what this special structure of equivalence relations means for the logic of knowledge.

Information models arise in many settings. They are used in philosophy as a representation of what thinking human agents know. Independently, they have been proposed in economics as a representation of what players know in the course of the game: the different worlds are then different stages in a play of the game, or even different "strategy profiles" describing precisely what each player is going to do throughout the game. Chapter 7 wil have more details on connections between logic and games. But information models have also been used for describing non-human agents in computer science, say, in describing different processors in a message-passing system for communication. There is a whole area called 'Agency' where epistemic logic plays a conspicuous role.

Semantics But first, we state how the epistemic language can be interpreted on our models. The format for this is like the truth definition that you have seen for the language

of predicate logic. We start by explaining when an atomic formula is true at a world, and then work our way up along all the constructions that build formulas:

Definition 5.24 (Truth Conditions for EL)

$$\begin{array}{rll} M, s \models p & \text{iff} & V \text{ makes } p \text{ true at } s \\ M, s \models \neg\varphi & \text{iff} & \text{not } M, s \models \varphi \\ M, s \models \varphi \vee \psi & \text{iff} & M, s \models \varphi \text{ or } M, s \models \psi \\ M, s \models \varphi \wedge \psi & \text{iff} & M, s \models \varphi \text{ and } M, s \models \psi \\ M, s \models \Box_i\varphi & \text{iff} & \text{for all } t \text{ with } s \to_i t\text{: } M, t \models \varphi \\ M, s \models \Diamond_i\varphi & \text{iff} & \text{for some } t \text{ with } s \to_i t \text{ it holds that } M, t \models \varphi. \end{array}$$

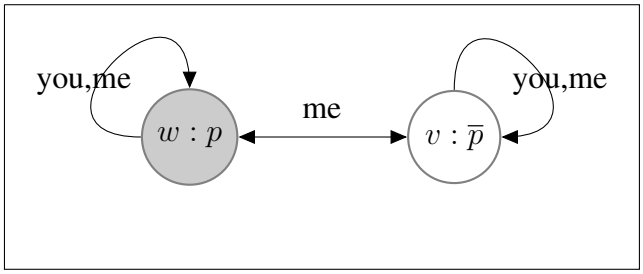
Here the clauses for the Boolean operations are exactly like those for propositional logic. The key clause is that for the knowledge operator $\Box_i\varphi$, which is read - and many people find this a helpful analogy - as a universal quantifier saying that φ is true in all accessible worlds. The epistemic operator $\Diamond_i\varphi$ can then be read dually as an existential quantifier saying that some accessible world exists satisfying φ. As was mentioned before, $\Diamond_i\varphi$ is equivalent to $\neg\Box_i\neg\varphi$. We will write

$$\Diamond\varphi$$

for this as an epistemic operator on its own, in cases where there is just a single agent (an intuitive reading would be that "φ is possible").

How does this work? Let us look at some examples.

Example 5.25 (Question and Answer Once More) Recall the above model:



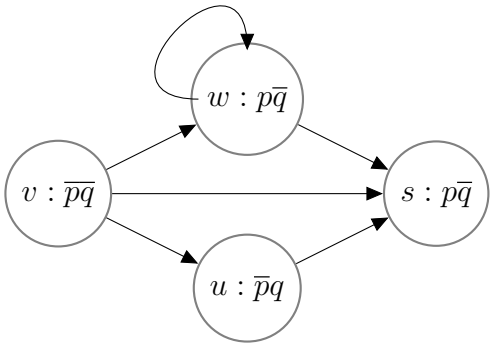
Intuitively, in the actual world (the shaded circle), I do not know whether p, but I know that you are fully informed about it. Spelling out the the above truth definition, one can check that this is right: in the actual world, $\neg\Box_{\text{me}}p \wedge \Box_{\text{me}}(\Box_{you}p \vee \Box_{you}\neg p)$ is true. For, you can see the following facts, listed in a table for convenience:

Formula	Worlds where true
p	w
$\neg p$	v
$\Box_{\text{you}} p$	w
$\Box_{\text{me}} p$	none
$\neg \Box_{\text{me}} p$	w, v
$\Box_{you} \neg p$	v
$\Box_{\text{me}} \neg p$	none
$\neg \Box_{\text{me}} \neg p$	w, v
$\neg \Box_{\text{me}} p \wedge \neg \Box_{\text{me}} \neg p$	w, v
$\Box_{you} p \vee \Box_{you} \neg p$	w, v
$\Box_{\text{me}} (\Box_{you} p \vee \Box_{you} \neg p)$	w, v.

You see the idea: evaluate simple formulas first, and find out in which worlds they are true. Then work your way upward to more complex formulas. In principle, this works no matter how large the model is, and how complex the formula you are evaluating.

Now let us also look at a much more abstract example, that does not admit an epistemic interpretation. This will allow you to see how our mechanism of interpretation works in the more general case.

Example 5.26 In the following graph, some worlds have a unique outgoing arrow to one accessible world, others have several, while there is also a world without any outgoing arrows at all: we will see in a moment what happens then.



The valuation is written into the diagram as before, marking worlds with proposition letters. Here are a few facts that you can easily check:

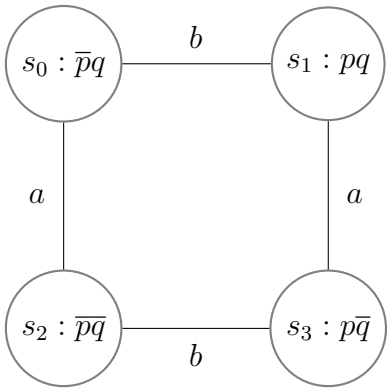
p	is true precisely in worlds	w, s
$\Box p$	is true precisely in worlds	w, u, s
$\Diamond \Box p$	is true precisely in worlds	v, w, u
$q \rightarrow p$	is true precisely in worlds	v, w, s
$\Box(q \rightarrow p)$	is true precisely in worlds	w, u, s

But we can also do something else. Each world in this model has something special, and we can describe this uniquely in the above language:

world w	is the only world satisfying	$p \wedge \Diamond p$
world v	is the only world satisfying	$\neg p \wedge \neg q$
world u	is the only world satisfying	q
world s	is the only world satisfying	$\neg\Diamond p$
world s	is the only world satisfying	$\Box\bot$.

In fact we could give many other formulas defining these four worlds uniquely. Note that $\Box\bot$ characterizes *endpoints* in the accessibility relation.

Example 5.27 (Interpreted Systems) An interpreted system is a process viewed as an information model. A process is a system that can move through various states; you will learn more about processes in Chapter 6. Our example consists of two sub-processes a and b. Propositional variable p describes the state of process a. If p is true, the state of a is 1, and if p is false the state of a is 0. Propositional variable q describes the state of process b, in a similar way. Both processes only know their own state. This corresponds to the following model.



As usual, a link labelled with a or b means that the connected states are indistinguishable for a or b, respectively. Call the model M. Now verify that $M, s_1 \models \Box_b q$ and that $M, s_3 \models \Box_a p$.

Exercise 5.28 Consider the process model from Example 5.27 again. We can describe that a knows its own state p by means of

$$(p \to \Box_a p) \wedge (\neg p \to \Box_a \neg p).$$

Similarly, b knows its own state q is expressed by:

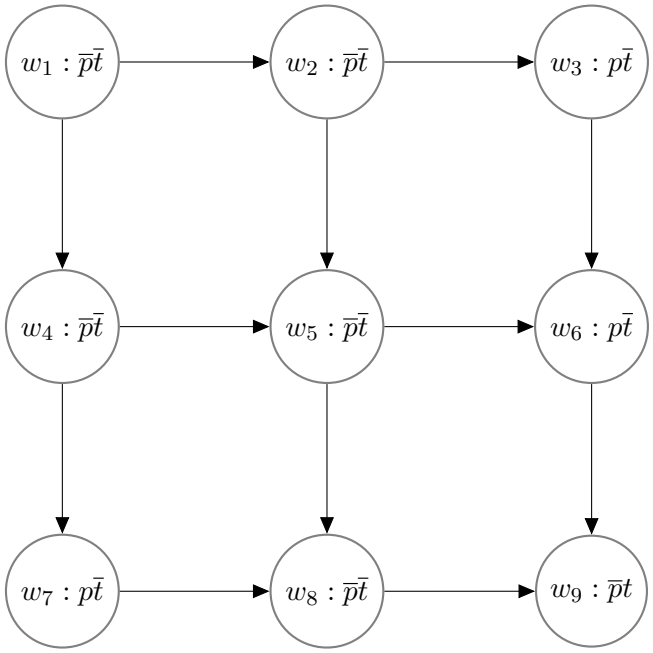
$$(q \to \Box_b q) \wedge (\neg q \to \Box_b \neg q).$$

Demonstrate that these two formulas hold in all states of the model.

Exercise 5.29 Consider the process model from Example 5.27 once more. Check that for any φ the following formula holds in all states of the model:

$$\Diamond_a \Box_b \varphi \to \Box_b \Diamond_a \varphi.$$

Exercise 5.30 (Treasure Island) Consider the following model with 9 states, and an accessibility relation allowing one step east or south (insofar as possible in the given picture) from each point. World w_9 satisfies the proposition letter t (the location of the 'treasure'), while pirates are standing at w_3, w_6, and w_7, marked by the proposition letter p.



In which worlds are the following epistemic formulas true?

(1) $\Diamond t$,

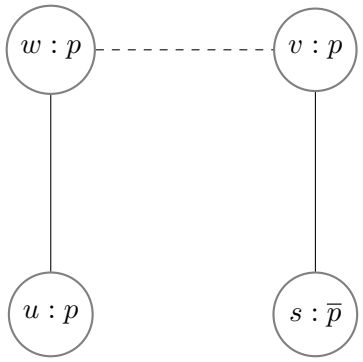
(2) $\Diamond \Box t$,

(3) $\Diamond p$,

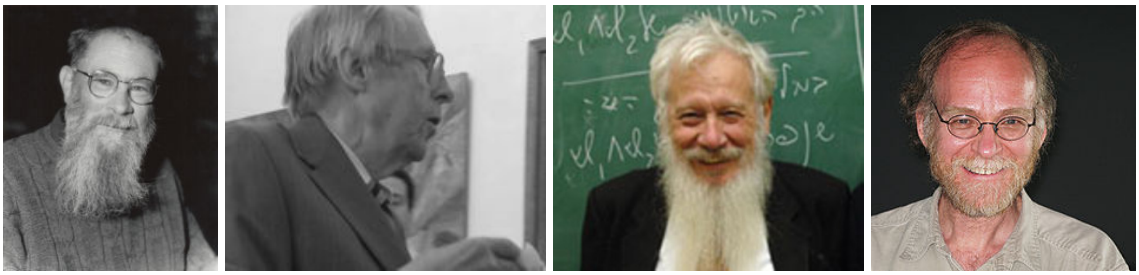
(4) $\Box \Diamond p$.

Next, for each world, find an epistemic formula which is only true at that state.

Exercise 5.31 Let's return to epistemic models in the proper sense, where the accessibility relations are all equivalences. Find epistemic formulas that uniquely define each world in the following model. The solid lines are links for agent 1, the dashed line is a link for agent 2.



Exercise 5.32 Consider the earlier models for the Three-cards scenario. Show in detail how the final model verifies the statements in the text about which player knows what about the cards, and the information of the other players.



In this picture gallery you see four people spanning the range of epistemic logic: David Lewis (a philosopher), Jaakko Hintikka (a philosopher/logician), Robert Aumann (an economist and Nobel prize winner), and Joe Halpern (a computer scientist).

5.6 Valid Consequence

We will now give definitions of valid formulas, and of valid consequence. First some examples of valid principles. We start with single operators on top of propositional logic.

$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi).$$

This is a kind of 'epistemic distribution': if one knows an implication then it follows that if one also knows the antecedent of the implication then one has to know the consequent too. We will adopt epistemic distribution in what follows, but we have to bear in mind that what it expresses is quite strong. It expresses the principle of so-called *logical omniscience*: our epistemic agents are perfect reasoners: they can draw the logical consequences from what they know.

$$\Diamond(\varphi \vee \psi) \leftrightarrow (\Diamond\varphi \vee \Diamond\psi).$$

Considering a disjunction possible is equivalent to holding at least one of the disjuncts for possible.

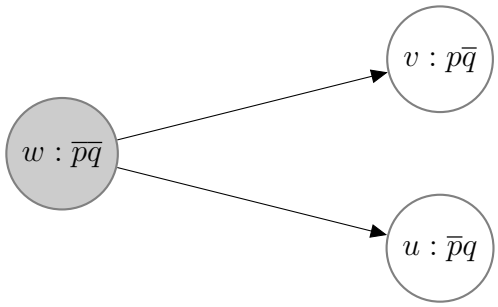
Next, there are the principles that express our abbreviation conventions:

$$\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi,$$

$$\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi.$$

A formula φ is called *invalid* if there is at least one model M with a world s where φ is false. Like in earlier chapters, such a pointed model (M, s) is called a counter-example for φ.

Example 5.33 (Counter-example for '$\Diamond$ over $\wedge$') In the actual world w in the following model M, $\Diamond p$ and $\Diamond q$ are true, but $\Diamond(p \wedge q)$ is not:



Here, one thing leads to another. Like in propositional and predicate logic, there are strong dualities between the two modalities and disjunction/conjunction, resulting in automatic further laws for $\Diamond$, $\Box$, $\wedge$, $\vee$. Thus, switching operators, the obvious valid counterpart to the distribution law $\Diamond(\varphi \vee \psi) \leftrightarrow (\Diamond\varphi \vee \Diamond\psi)$ is the following principle:

$$\Box(\varphi \wedge \psi) \leftrightarrow (\Box\varphi \wedge \Box\psi).$$

On the same pattern of standing and falling together, typically *invalid* principles are:

$$\Diamond(\varphi \wedge \psi) \leftrightarrow (\Diamond\varphi \wedge \Diamond\psi),$$

$$\Box(\varphi \vee \psi) \leftrightarrow (\Box\varphi \vee \Box\psi).$$

To see that $\Box(\varphi \vee \psi) \rightarrow (\Box\varphi \vee \Box\psi)$ is invalid, look at the special case where ψ equals $\neg\varphi$. As a tautology, $\varphi \vee \neg\varphi$ has to be true in any situation. Then it also has to hold that $\Box(\varphi \vee \neg\varphi)$ is true anywhere, for all tautologies are known. But it does certainly not follow that either $\Box\varphi$ or $\Box\neg\varphi$, for φ might well express some fact about which nothing is known. A concrete counterexample is the model above, where $\Box(p \vee \neg p)$ is true in world w, but both $\Box p$ and $\Box\neg p$ are false in w.

Correspondence between special relational properties and epistemic axioms Are there also interesting laws that express epistemic intuitions? That depends. First, consider one single agent. Here are three well-known axioms with prominent epistemic interpretations:

Veridicality $\Box\varphi \to \varphi$.

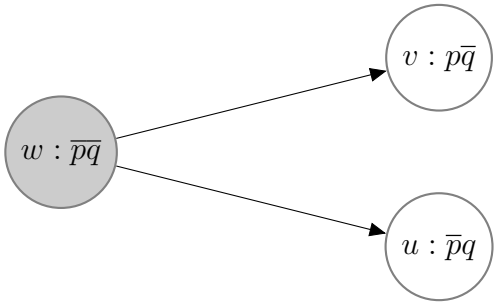
Positive Introspection $\Box\varphi \to \Box\Box\varphi$.

Negative Introspection $\neg\Box\varphi \to \Box\neg\Box\varphi$.

The first of these seems uncontroversial: knowledge has to be 'in sync' with reality, or it does not deserve to be called knowledge. If something you were firmly convinced of turns out to be false, then you may have *thought* you knew it to be true, while in actual fact you did not know it. But the other two have been much discussed. Positive introspection says that agents know what they know, and negative introspection that agents know what they do not know, and both principles seem rather strong. They seem debatable, for they assume that our epistemic agents, in addition to their logical omniscience in terms of powers of inference (encoded in the earlier distribution axiom), they now also have capacities of unlimited introspection into their own epistemic states.

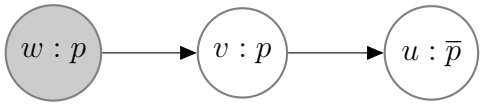
Formally, these axioms are not valid in the most general semantics that we have given.

Example 5.34 Consider the model we gave above.



Notice that in the actual world $p \vee q$ does not hold, but $\Box(p \vee q)$ does. Thus, in the actual world of the model we have $\neg(\Box(p \vee q) \to (p \vee q))$. In other words, veridicality does not hold in the model, which shows that this is not an appropriate model for representing knowledge.

Example 5.35 Consider the following model.



Note that in the actual world of the model $\Box p$ and $\neg\Box\Box p$ are both true. Thus, positive introspection does not hold in the model.

Exercise 5.36 Find a simple example of a model where $\neg\Box p \to \Box\neg\Box p$ is not true in the actual world.

It is customary to study information models with certain restrictions on their accessibility relations. Recall the "equivalence relations" that we have mentioned briefly before. If we demand that our accessibility relations in information models are of this special kind, then the above axioms become valid. In fact, the validity of each of the above axioms there is some feature in the notion of an equivalence relation that is crucially involved. The following table show this (using R for the epistemic accessibility relation):

Name	EL formula	relational principle	PL formula
Veridicality	$\Box\varphi \to \varphi$	reflexivity	$\forall x Rxx$
Pos Introsp	$\Box\varphi \to \Box\Box\varphi$	transitivity	$\forall x\forall y\forall z((Rxy \wedge Ryz) \to Rxz)$
Neg Introsp	$\neg\Box\varphi \to \Box\neg\Box\varphi$	euclidicity	$\forall x\forall y\forall z((Rxy \wedge Rxz) \to Ryz)$.

When equivalence relations were introduced on page 5-16, you saw *symmetry* instead of *euclidicity*. The following two exercises explain the connection.

Exercise 5.37 Show that every transitive and symmetric relation R is euclidic.

Exercise 5.38 Show that every euclidic and reflexive relation R is symmetric.

It follows from these two exercises that every equivalence relation is euclidic, and that every relation that is reflexive, transitive and euclidic is an equivalence relation.

To explain the connection between $\Box\varphi \to \varphi$ and reflexivity, first note that if $M, s \models \Box\varphi$ and it is given that the relation for $\Box$ is reflexive, $M, s \models \varphi$ has to hold as well, as s is among the worlds that are accessible from s. But this means that $\Box\varphi \to \varphi$ holds in all reflexive models.

To connect positive introspection and transitivity, notice that the positive introspection principle $\Box\varphi \to \Box\Box\varphi$ is equivalent to $\Diamond\Diamond\varphi \to \Diamond\varphi$ (use the definition of $\Diamond$, and do contraposition, replacing φ by $\neg\varphi$ and cancelling double negations).

Suppose $M, s \models \Diamond\Diamond\varphi$, and assume that it is given that the accessibility relation of M is transitive. Then by the truth definition there is a world t with Rst and $M, t \models \Diamond\varphi$. So, again by the truth definition, there is a world u with Rtu and $M, u \models \varphi$. Because R is transitive it follows from Rst and Rtu that Rsu, and therefore by the truth definition, $M, s \models \Diamond\varphi$. Since s was arbitrary, it follows that $\Diamond\Diamond\varphi \to \Diamond\varphi$ is valid on M.

Exercise 5.39 Which one of the following two implications is valid on models with equivalence relations? Draw a counter-example for the other:

(1) $\Diamond_1\Box_2\varphi \to \Diamond_2\Diamond_1\varphi$.

(2) $\Diamond_1\Box_2\varphi \to \Diamond_2\Box_1\varphi$.

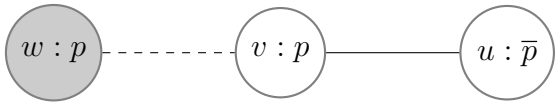
Exercise 5.40 Explain why the axiom $\neg\Box\varphi \to \Box\neg\Box\varphi$ is valid on every information model with an accessibility relation that satisfies $\forall x\forall y\forall z((Rxy \wedge Rxz) \to Ryz)$.

What about iterating knowledge of different agents? Consider $\Box_1\Box_2\varphi$. Saying that the two knowledge operators *commute* would boil down to: "I know what you know if and only if you know what I know." That does not sound plausible at all. And indeed, even when we impose the above special validities for single agents by using equivalence relations, our logical system will not validate laws that perform significant changes between knowledge of different agents. There can be interesting informative patterns on what agents know about what other agents know, to be sure, but such patterns always emerge in specific communicative contexts, as effects of what goes on in the communication. More on this below. The following example gives a counterexample to commutation of knowledge of several agents:

Example 5.41 (Your knowledge and mine do not commute) The following model is a counter-example to the putative implication

$$\Box_1\Box_2 p \to \Box_2\Box_1 p.$$

Its antecedent is true in the actual world, but its consequent is false (assume the solid line pictures the accessibility for agent 1, and the dashed line the accessibility for agent 2). Note that arrowheads and reflexive arrows are omitted from the example, since we assume that the accessibility relations are equivalences.



Such implications only hold when agents stand in special informational relationships. For example, $\Box_1\Box_2\varphi \to \Box_2\Box_1\varphi$ would hold in case it is given that $R_2 \circ R_1 \subseteq R_1 \circ R_2$.

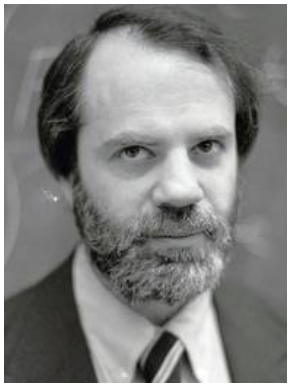
5.7 Proof

As we have done for propositional and predicate logic, in epistemic logic, too, we can establish validity by means of proof. Here are a few basic principles of epistemic reasoning in this format. Our main point is to show how the earlier style of thinking applies here too, even though our topic of information and knowledge seems quite different from that of Chapters 2 and 4. One more reason for pursuing this is that much of the pure and applied research on validity in epistemic logic, and the dynamic logic of action to follow

in Chapter 6, has focused on axiomatic proof systems, so understanding them is your key to the literature.

We will start out with a proof system for any logic with accessibilities (any so-called *modal logic*), and we will then enrich the system with special axioms describing agents with special powers for which one can derive more laws. Practically, this means that your reasoning gets richer when you know that you are dealing with agents having, e.g., the power of introspection.

The proof system we start out with is called the minimal modal logic, or the modal logic K. The name K is a historical relict; it refers to the inventor of the 'possible world' semantics for modal logic, Saul Kripke (b. 1940), and it has nothing to do with the 'K' of 'Knowledge'.



Saul Kripke

Definition 5.42 (Proof system for the minimal modal logic K) The proof system for the minimal modal logic K is given by:

(1) All propositional tautologies are theorems.

(2) All formulas of the form $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ are theorems. This is called the **distribution axiom schema** or the **K axiom schema**.

(3) If φ and $\varphi \to \psi$ are theorems, then ψ is a theorem. This is the familiar **modus ponens rule**.

(4) If φ is a theorem, then $\Box\varphi$ is also a theorem. This rule is called the **necessitation rule**.

This system is an extension of the proof system for propositional logic, and it is usual, in working with this system, to perform any propositional reasoning steps without spelling out details.

We will use $\vdash \varphi$ for "φ is provable" or "φ is a theorem". E.g., $\vdash \Box p \vee \neg\Box p$, because $\Box p \vee \neg\Box p$ has the form $\varphi \vee \neg\varphi$ of a propositional tautology.

Example 5.43 (Distribution rules, 1) If $\vdash \varphi \to \psi$, then $\vdash \Box\varphi \to \Box\psi$.

(1) $\vdash \varphi \to \psi$ assumption

(2) $\vdash \Box(\varphi \to \psi)$ necessitation rule on 1

(3) $\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ distribution axiom

(4) $\vdash \Box\varphi \to \Box\psi$ modus ponens on 2, 3

Example 5.44 (Distribution rules, 2) If $\vdash \varphi \to \psi$, then $\vdash \Diamond\varphi \to \Diamond\psi$.

(1) $\vdash \varphi \to \psi$ assumption

(2) $\vdash \neg\psi \to \neg\varphi$ propositional logic, 1

(3) $\vdash \Box\neg\psi \to \Box\neg\varphi$ by subroutine of previous example

(4) $\vdash \neg\Box\neg\varphi \to \neg\Box\neg\psi$ propositional logic, 3

(5) $\vdash \Diamond\varphi \to \Diamond\psi$ definition of $\Diamond$.

Example 5.45 Formal proof of $\vdash \Box(\varphi \wedge \psi) \leftrightarrow (\Box\varphi \wedge \Box\psi)$. We will do this in two steps, proving first the implication from left to right and then the implication from right to left. Putting the two together is then just a matter of applying a propositional reasoning step.

(1) $\vdash (\varphi \wedge \psi) \to \varphi$ propositional tautology

(2) $\vdash \Box(\varphi \wedge \psi) \to \Box\varphi$ $\Box$ distribution, example 5.43

(3) $\vdash \Box(\varphi \wedge \psi) \to \Box\psi$ similarly, starting from $(\varphi \wedge \psi) \to \psi$

In propositional logic we can infer from $\vdash \varphi \to \psi$ and $\vdash \varphi \to \chi$ that $\vdash \varphi \to (\psi \wedge \chi)$. In the present system we can reason in the same way. So we get from the above:

(4) $\vdash \Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$.

Now for the other direction:

(5) $\vdash \varphi \to (\psi \to (\varphi \wedge \psi))$ propositional tautology

(6) $\vdash \Box(\varphi \to (\psi \to (\varphi \wedge \psi)))$ necessitation, 5

(7) $\vdash \Box\varphi \to \Box(\psi \to (\varphi \wedge \psi))$ distribution axiom, prop logic, 6

(8) $\vdash \Box\varphi \to (\Box\psi \to \Box(\varphi \wedge \psi))$ distribution axiom, prop logic, 7

(9) $\vdash (\Box\varphi \wedge \Box\psi) \to \Box(\varphi \wedge \psi)$ prop logic, 8

Finally, we use propositional logic to put 5 and 9 together, and we get:

$$\vdash \Box(\varphi \wedge \psi) \leftrightarrow (\Box\varphi \wedge \Box\psi).$$

Exercise 5.46 Prove the following:

$$\vdash (\Diamond\varphi \wedge \Box(\varphi \rightarrow \psi)) \rightarrow \Diamond\psi.$$

Stronger modal logics increase deductive power by adding further axiom schemata to the minimal logic K. Then the flavour of finding derivations may change, as you develop a feeling for what the additional power gives you. Here are some well-known examples:

Definition 5.47 (T, S4, and S5) The modal logic T arises from K by adding the axiom schema of Veridicality $\Box\varphi \rightarrow \varphi$. The logic S4 adds the schema $\Box\varphi \rightarrow \Box\Box\varphi$ to T (we have encountered this above as the principle of Positive Introspection). Finally, the logic S5 adds the schema $\Diamond\varphi \rightarrow \Box\Diamond\varphi$ to S4 (encountered above as the principle of Negative Introspection).

Example 5.48 (Alternative definition of S5) This example shows that S5 also arises from K by adding the axiom schemes of Veridicality $\Box\varphi \rightarrow \varphi$ and $\Diamond\varphi \rightarrow \Box\Diamond\varphi$ (the principle of Negative Introspection). To show that, we derive the principle of Positive Introspection from $\Box\varphi \rightarrow \varphi$ and $\Diamond\varphi \rightarrow \Box\Diamond\varphi$.

(1) $\vdash \Diamond\neg\varphi \rightarrow \Box\Diamond\neg\varphi$ negative introspection axiom schema

(2) $\vdash \Diamond\Box\varphi \rightarrow \Box\varphi$ prop logic, $\Diamond$ def, 1

(3) $\vdash \Box(\Diamond\Box\varphi \rightarrow \Box\varphi)$ necessitation, 2

(4) $\vdash \Box\Diamond\Box\varphi \rightarrow \Box\Box\varphi$ distribution, prop logic, 3

(5) $\vdash \Diamond\Box\varphi \rightarrow \Box\Diamond\Box\varphi$ instance of schema $\Diamond\varphi \rightarrow \Box\Diamond\varphi$

(6) $\vdash \Diamond\Box\varphi \rightarrow \Box\Box\varphi$ prop logic, 4, 5

(7) $\vdash \Box\Diamond\neg\varphi \rightarrow \Diamond\neg\varphi$ instance of T schema

(8) $\vdash \Box\varphi \rightarrow \Diamond\Box\varphi$ prop logic, 7

(9) $\vdash \Box\varphi \rightarrow \Box\Box\varphi$ prop logic, 8, 6.

Example 5.49 (Yet another definition of S5) The system that results by adding the schemes for veridiality, positive introspection, plus the following principle of symmetry $\varphi \rightarrow \Box\Diamond\varphi$ to K is also S5. We prove that the principle of negative introspection follows from $\varphi \rightarrow \Box\Diamond\varphi$ and $\Box\varphi \rightarrow \Box\Box\varphi$:

(1) $\vdash \Diamond\varphi \rightarrow \Box\Diamond\Diamond\varphi$ instance of $\varphi \rightarrow \Box\Diamond\varphi$

(2) $\vdash \Diamond\Diamond\varphi \rightarrow \Diamond\varphi$ prop logic, from $\Box\neg\varphi \rightarrow \Box\Box\neg\varphi$

(3) $\vdash \Box(\Diamond\Diamond\varphi \rightarrow \Diamond\varphi)$ necessitation, 2

(4) $\vdash \Box\Diamond\Diamond\varphi \rightarrow \Box\Diamond\varphi$ distribution, prop logic, 3

(5) $\vdash \Diamond\varphi \rightarrow \Box\Diamond\varphi$ prop logic, 1, 4

And here is a derivation of $\varphi \rightarrow \Box\Diamond\varphi$, using only the K schema, the T schema and the schema of negative introspection:

(1) $\vdash \Box\neg\varphi \rightarrow \neg\varphi$ instance of $\Box\varphi \rightarrow \varphi$

(2) $\vdash \varphi \rightarrow \Diamond\varphi$ def of $\Diamond$, prop logic, 1

(3) $\vdash \Diamond\varphi \rightarrow \Box\Diamond\varphi$ axiom

(4) $\vdash \varphi \rightarrow \Box\Diamond\varphi$ prop logic, 2, 3

Example 5.50 (Reduction of Modalities for S5) We will now prove that S5 allows reduction of modalities, where a modality is a list of modal operators $\Box, \Diamond$. First we show that $\Diamond\varphi \leftrightarrow \Box\Diamond\varphi$ is a theorem:

(1) $\vdash \Diamond\varphi \rightarrow \Box\Diamond\varphi$ axiom schema of Negative Introspection

(2) $\vdash \Box\Diamond\varphi \rightarrow \Diamond\varphi$ instance of T axiom schema

(3) $\vdash \Diamond\varphi \leftrightarrow \Box\Diamond\varphi$ prop logic, 1, 2

By contraposition and the interdefinability of $\Box$ and $\Diamond$ we get from this that $\Box\varphi \leftrightarrow \Diamond\Box\varphi$ is also a theorem. It is easy to show that $\Box\varphi \leftrightarrow \Box\Box\varphi$ is an S5 theorem:

(1) $\vdash \Box\Box\varphi \rightarrow \Box\varphi$ instance of T axiom schema

(2) $\vdash \Box\varphi \rightarrow \Box\Box\varphi$ positive introspection schema

(3) $\vdash \Box\varphi \leftrightarrow \Box\Box\varphi$ prop logic, 1, 2

By contraposition and the interdefinability of $\Box$ and $\Diamond$ we get from this that $\vdash \Diamond\varphi \leftrightarrow \Diamond\Diamond\varphi$.

Thus we see, that in S5, it is always the innermost operator that counts: $\Diamond\Box$ and $\Box\Box$ reduce to $\Box$, $\Box\Diamond$ and $\Diamond\Diamond$ reduce to $\Diamond$.

But note that we are talking here about a system for reasoning about a single agent. In the multi-agent logic of knowledge, $\Box_i\Box_j\varphi$ does *not* reduce to a single modality.

Exercise 5.51 Which of the following two implications is valid? Give an informal argument, and also a formal proof in the minimal logic K:

(1) $\Box(p \to q) \to (\Diamond p \to \Diamond q)$.

(2) $(\Diamond p \to \Diamond q) \to \Box(p \to q)$.

As for the invalid formula, give a counterexample (draw the model).

Exercise 5.52 Prove the formula $\Box(p \vee q) \to (\Diamond p \vee \Box q)$.

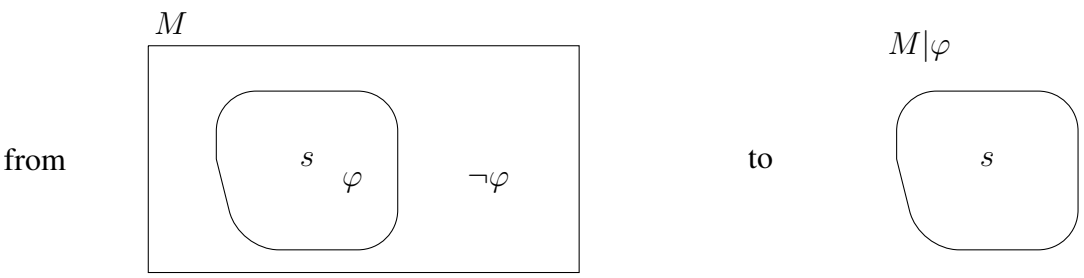
5.8 Information Update

In Section 5.3 we explained the intuitive notion of information update through successive elimination of possibilities. We will now make this more precise, and explain how information sources like observation and communication through update can be dealt with in a formal way, by describing the process by which information states are changed by incoming information, to yield new information states (or: 'updated' information states).

This update process works over pointed information models (M, s) with s the actual world. When new information arrives, we can think of this as a public announcement of some true proposition φ, where "true" means that φ holds in the actual world: $M, s \models \varphi$. 'Public announcement' is a generic technical term for various things that could be going on: it could a public broadcast, but it might just as well be some publicly observable happening like the outbreak of a thunderstorm, or a sunset witnessed by all.

Definition 5.53 (Updating via definable submodels) For any epistemic model M, world s, and formula φ true at s, the model $(M|\varphi, s)$ (M relativized to φ at s) is the sub-model of M whose domain is the set $\{t \in W_M \mid M, t \models \varphi\}$ (where W_M indicates the set of worlds of M).

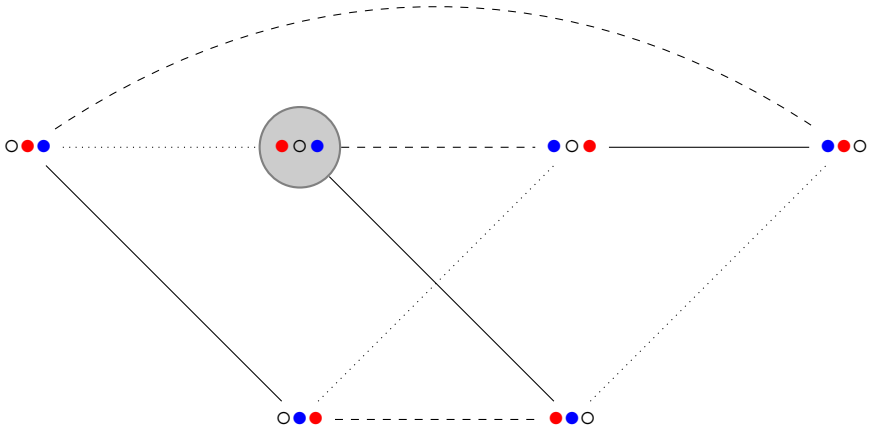
Drawn in a simple picture, such an update step goes



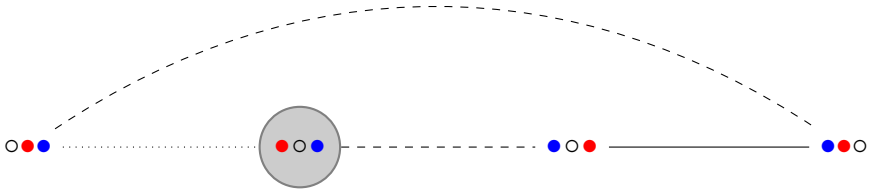
The fact that we eliminate means that the $\neg\varphi$-worlds are ruthlessly discarded: they are gone forever. This is often called 'hard information', an act which changes the current

model irrevocably as a response to some totally trusted source. You can think of this as a typical step of communication when something is said by a trusted authority, but another good reading is as an act of public observation, whether or not stated in language.

Example 5.54 We return to the Three Cards example (Example 5.6), to see how this fits this format. Here is the picture of the situation after the card deal again:



The question of agent 2 to agent 1 "Do you have the blue card", if honest, implies the statement "I know that I do not have the blue card", or, formally, $\Box_2 \neg b_2$ (if we assume a language with proposition letters b_1, b_2, b_3 with the obvious meanings). From $\Box_2 \neg b_2$ it follows that $\neg b_2$; the worlds where $\neg b_2$ is true are precisely the worlds ○●●, ●○●, ●●○ and ●●○, so the model restricted to $\neg b_2$ is indeed as we saw before:

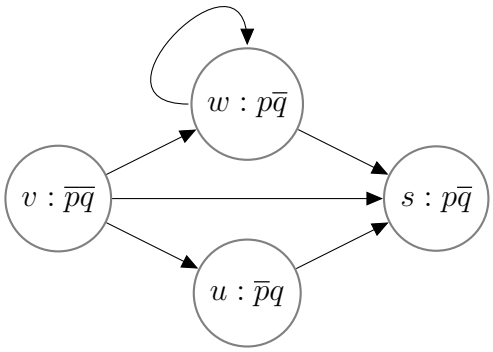


One word about the term "public": we use this, because the above elimination operation is the same for all the agents that live in the model. They all "see" the same structure afterwards. In reality, of course, the world is full of differences in observation, and even hiding of information. The information flow in scenarios with privacy is much more subtle than what we discuss here: see Outlook Section 5.10 below for a few thoughts.

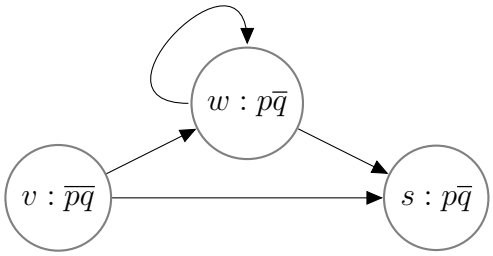
These diagrams of a jump from one model to another seem to be simple, but they are trickier than you might think. Crucially, truth values of epistemic formulas at a world may change in an update step as depicted here. For instance, when a public announcement $!p$ is made, only the p-worlds remain (we use $!p$ for the public announcement of the fact p, and

more generally, $!\varphi$ for the public announcement of φ). But at p-worlds where formerly, the agent did not know that p because of the presence of some accessible $\neg p$-world, the agent knows p after the update, since there are only p-worlds around now.

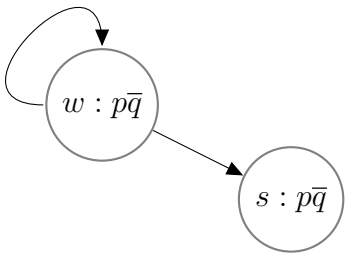
Example 5.55 Recall the model of example 5.26, repeated here:



If this model is called M, what is $M|p \vee \Diamond q$? To calculate the model, just check where $p \vee \Diamond q$ is true, and we get the set $\{v, w, s\}$ (for u is the only world where $p \vee \Diamond q$ fails. Therefore, $M|p \vee \Diamond q$ looks like this:



Now we can evaluate $p \vee \Diamond q$ once again. This time v fails the formula, because the world that made $\Diamond q$ true before is no longer present. We get:



Finally, if we update with $p \vee \Diamond q$ yet another time, nothing changes anymore.

Exercise 5.56 Find a simple example of a model M with the property that M, $M|\Diamond p$, $(M|\Diamond p)|\Diamond p$, and $((M|\Diamond p)|\Diamond p)|\Diamond p$ are all different.

This update mechanism, simple though it may seem, explains many knowledge puzzles, one of which is an evergreen, as it packs many relevant topics into one simple story.[1]

Example 5.57 (Muddy Children) Three children play outside, and two of them get mud on their foreheads. They can only see the other children's foreheads, so they do not know whether they themselves are muddy or not. (This is an inverse of our card games.) Now their father says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" (with 'knowing if you are dirty' the father means 'knowing that you are dirty if you are and knowing that you are not dirty if you are not'). The children answer truthfully. As questions and answers repeat, what happens? Assume that the children are all perfect reasoners, and this is commonly known among them.

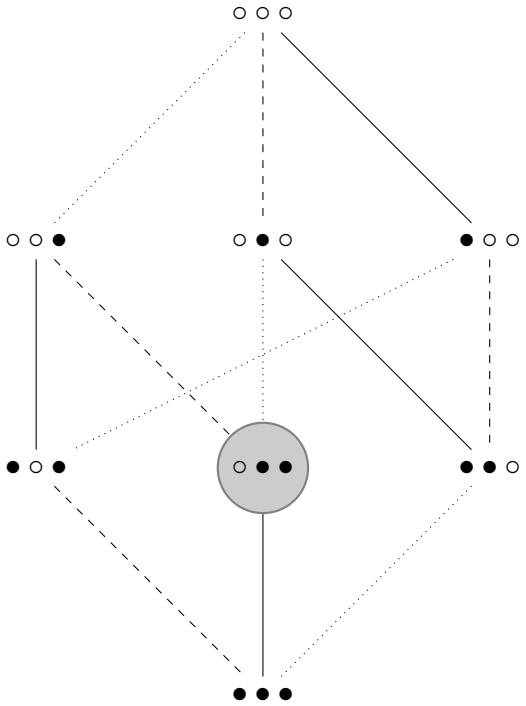
Nobody knows in the first round. But in the next round, each muddy child can reason like this: "Suppose I were clean. Then the one dirty child I see would have seen only clean children, and so she would have known that she was dirty at once. But she did not. So I must be dirty, too!"

So what happens is this:

	1	2	3
"At least one of you is dirty"	○	●	●
"Does anyone know if he is dirty?"	"No"	"No"	"No"
"Does anyone know if he is dirty?"	"No"	"Yes"	"Yes"
"Do you know if you are dirty?"	"Yes"		

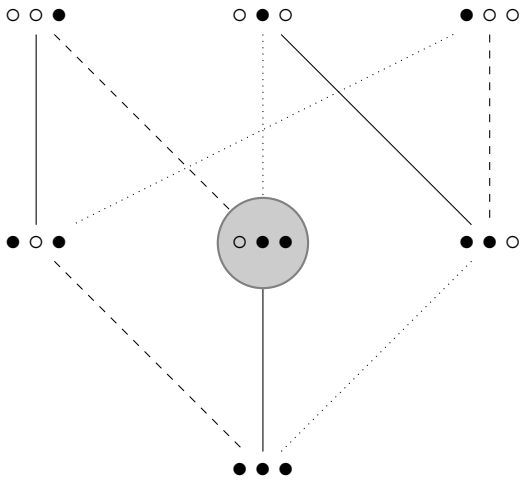
In the initial model, eight possible worlds assign "dirty" or "clean" to each child. We indicate this pictorially as ○ ● ●, for "child 1 clean, children 2 and 3 dirty" and so on. Suppose the actual situation is ○ ● ●. In the language, we assume we have proposition letters d_1, d_2, d_3, for "child 1 is dirty", "child 2 is dirty", "child 3 is dirty", respectively. So the situation ○ ● ● is described by $\neg d_1 \wedge d_2 \wedge d_3$. We can represent the initial model like this:

[1] The Muddy Children Puzzle has been doing the rounds for a long time. The currently earliest known source is a long footnote in a German 1832 translation of the French novel sequence *La vie de Gargantua et de Pantagruel* by Rabelais. This is a version with charcoal on noses, instead of mud on foreheads.



Solid lines are links for child 1, dashed lines are links for child 2, dotted lines links for child 3. There is a solid line between ○ ○ ○ and ● ○ ○, for the first child cannot distinguish the situation where the are all clean from the situation where the other two are clean and she is dirty, and so on. So the fact that the children all know about the others' faces, and not their own, is reflected in the accessibility links in the diagram.

Now let us see what happens when father makes his announcement. In a formula: $d_1 \vee d_2 \vee d_3$. The effect of this is that the world with ○ ○ ○ disappears:

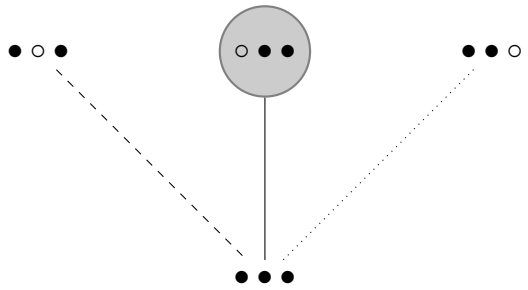


Now in response to father's first question "Does any of you know whether you are dirty?"

the children all say "No". These are three announcements:

$$\neg\Box_1 d_1 \wedge \neg\Box_1\neg d_1, \neg\Box_2 d_2 \wedge \neg\Box_2\neg d_2, \neg\Box_3 d_3 \wedge \neg\Box_3\neg d_3.$$

The first of these is false in world • ○ ○, the second is false in world ○ • ○, and the third is false in world ○ ○ •. (But note that these three formulas would all have been true everywhere in the initial model, before father's announcement.) So an update with these announcements results in the following new model:



Now father asks again, and the answers are:

$$\neg\Box_1 d_1 \wedge \neg\Box_1\neg d_1, \Box_2 d_2 \vee \Box_2\neg d_2, \Box_3 d_3 \vee \Box_3\neg d_3.$$

These answers make • ○ •, • • ○ and • • • disappear, and the final update results in a situation where only the actual world remains:

In this sequence of models, domains decrease in size stepwise: from 8 worlds to 7, then to 4, then down to 1. With k muddy children, k rounds of the simultaneous assertion "I do not know my status" yield common knowledge of which children are dirty. A few additional assertions by those who now know are then enough to establish common knowledge of the complete distribution of the mud on the faces of the group.

Exercise 5.58 Suppose that in the Three-Card example, the question of Player 2 is not treated as informative. What happens then? Draw the two updates.

Exercise 5.59 Suppose that in our Muddy Children scenario with 3 kids, the children speak in turn, each time starting from the first child, then the second, and finally the third. What happens? Draw the successive diagrams.

Exercise 5.60 Suppose in the Muddy Children scenario, with 3 children, the father says "At least one of you is clean", and then the same procedure is followed as before. Compute the update sequence, and explain what happens.

Exercise 5.61 Three men are standing on a ladder, each wearing a hat. Each can see the colours of the hats of the people below him, but not his own or those higher up. It is common knowledge that only the colours red and white occur, and that there are more white hats than red ones. The actual order is white, red, white from top to bottom. Draw the information model. The top person says: "I know the color of my hat". Is that true? Draw the update. Who else knows his color now? If that person announces that he knows his colour, what does the bottom person learn?

We end with some examples with a more philosophical flavour. One of the methods that philosophers have developed to clear up conceptual muddles is the use of bits of formal proofs to make their arguments precise. This is one more motivation for us to teach you at least some basic skills in proof reading, and proof search.

Example 5.62 (Moore-type sentences) Public announcement of atomic facts p (or more generally, purely propositional facts) makes them common knowledge. But not all events $!\varphi$ result in common knowledge of φ. A counter-example are so-called 'Moore-type' sentences. In a question-answer scenario, let the answerer A say truly

$$p \wedge \neg \Box_Q p \quad \text{"}p\text{, but you don't know it"}$$

This removes Qs ignorance about p, and thus makes itself false: true sentences like this lead to knowledge of their negation. This also occurred with the Muddy Children, where the last assertion of ignorance led to knowledge.

Example 5.63 (Verificationism and the Fitch paradox) The general verificationist thesis (VT) says that what is true can be known – or formally:

$$\varphi \to \Diamond \Box \varphi. \tag{VT}$$

A surprising argument by the philosopher Frederic Fitch (1908–1987) trivializes this principle, taking the substitution instance

$$(\varphi \wedge \neg \Box \varphi) \to \Diamond \Box (\varphi \wedge \neg \Box \varphi).$$

Then we have the following chain of three conditionals (say, in the weak modal logic T):

(1) $\Diamond \Box (\varphi \wedge \neg \Box \varphi) \to \Diamond (\Box \varphi \wedge \Box \neg \Box \varphi)$

(2) $\Diamond (\Box \varphi \wedge \Box \neg \Box \varphi) \to \Diamond (\Box \varphi \wedge \neg \Box \varphi)$

(3) $\Diamond (\Box \varphi \wedge \neg \Box \varphi) \to \bot$.

Here, $\bot$ is shorthand for a formula that is always false, say $p \wedge \neg p$.

Thus, a contradiction has been derived from the assumption $\varphi \wedge \neg \Box \varphi$, and we have shown over-all that φ implies $\Box \varphi$, making truth and knowledge equivalent. Here it seems plausible to read the modality as an event of getting hard information, and then the point is again that the Moore sentence $\varphi \wedge \neg \Box \varphi$ cannot be truly announced without making itself false.

5.9 The Logic of Public Announcement

In Chapters 3, 2, 4 we did not have to worry about the subtleties of information change, since each model pictured just one single unchanging setting. Now we have added change, and a logical system helps us be precise and consistent about reasoning in the presence of change. To do so, we must bring the informational actions themselves explicitly into the logic.

Language A suitable language for this is a combination of epistemic and dynamic logic. Dynamic logic is a tool for studying action – much more about this in Chapter 6 – and languages for dynamic logic include action expressions. We will focus here on the action of making public announcements.

Definition 5.64 (Language and semantics of public announcement) The language of public announcement logic PAL (without common knowledge) is the epistemic language with added action expressions, as well as dynamic modalities for these, defined by the syntax rules:

$$\begin{array}{rrl} \text{Formulas } \varphi & ::= & p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid (\varphi \wedge \varphi) \mid \Box_i\varphi \mid \Diamond_i\varphi \mid [A]\varphi \mid \langle A\rangle\varphi \\ \text{Action expressions } A & ::= & !\varphi \end{array}$$

In Outlook Section 5.11 we will extend this language with an operator for expressing common knowledge.

Note the mutual recursion in this definition: action expressions contain formulas which may contain action expressions which contain formulas, and so on. There is no harm in this, for each embedded action is syntactically less complex than the expression it is a part of, and the recursion comes to an end.

Semantics The epistemic language is interpreted as before in Section 5.5, while the semantic clause for the new dynamic action modality is 'forward-looking' among models as follows:

$$\begin{array}{lll} M, s \models [!\varphi]\psi & \text{iff} & M, s \models \varphi \text{ implies } M|\varphi, s \models \psi \\ M, s \models \langle !\varphi\rangle\psi & \text{iff} & M, s \models \varphi \text{ and } M|\varphi, s \models \psi. \end{array}$$

This language can make typical assertions about knowledge change like $[!\varphi]\Box_i\psi$, which states what an agent i will know after having received the hard information that φ.

Example 5.65 We return again to example 5.6 (and 5.54). We evaluate the following formula in the model of the initial situation, just after the card deal:

$$[!\neg b_2][!\neg b_1]\Box_1(\neg K_3 w_1 \wedge \neg K_3 r_1).$$

What this says is: after the public announcement "2 does not have blue", followed by the public announcement update with "1 does not have blue", player 1 knows that player 3 does not know 1's card. To compute whether this is true, just perform the updates and evaluate $\Box_1(\neg K_3 w_1 \wedge \neg K_3 r_1)$ in the actual world of the resulting model:

The formula is true in the actual world. So the original formula is true in the actual world of the model we started out with.

Exploring updates Call a public announcement $!\varphi$ *factual* if φ is a purely propositional formula (no modality $\Box_i$ or $\Diamond_i$ occurs in φ). Suppose φ is true in the actual world. Then an announcement of φ makes φ known to all agents. In fact, φ becomes common knowledge as a result of its announcement. For epistemic formulas this need not be the case. Indeed, we have seen some examples of that already: Moore-type formulas, but also the public announcements of ignorance in the Muddy Children scenario, where announcing ignorance may create knowledge. A formula like $\neg\Box p$ may become known after being announced, but it need not be. See the next exercise.

Exercise 5.66 Find a pointed model (M, s) where $M|\neg\Box p \models \neg\Box p$, and a pointed model (N, t) where $N|\neg\Box p \models \Box p$.

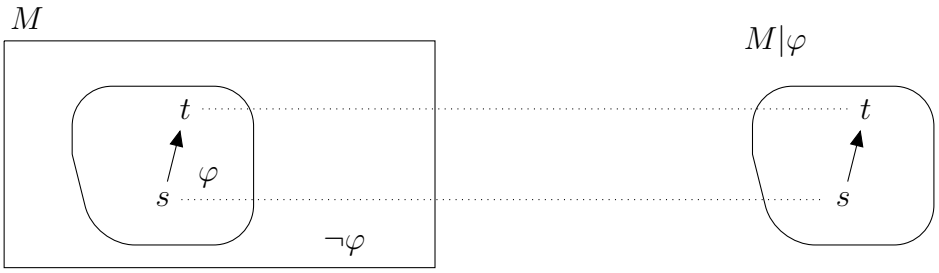
And we have seen that a Moore sentence like $(p \wedge \neg\Box p)$ always makes itself false when announced. Epistemic formulas that make themselves false when announced are not oddities but can be very useful: see the ignorance assertions in Muddy Children above, that led to knowledge.

Valid reasoning and proof Reasoning about information flow in public update satisfies axioms, just as we saw before with reasoning about knowledge and information at some fixed stage, i.e., some given information model. This reasoning revolves around the following dynamic 'recursion equation', which relates new knowledge to old knowledge the agent had before:

Valid PAL equivalence, useful for axiomatizing announcements The following equivalence is valid for PAL:

$$[!\varphi]\Box_i\psi \leftrightarrow (\varphi \rightarrow \Box_i(\varphi \rightarrow [!\varphi]\psi)).$$

To see why this is true, compare the two models (M, s) and $(M|\varphi, s)$ before and after the update.



The formula $[!\varphi]\Box_i\psi$ says that, in $M|\varphi$, all worlds t that are i-accessible from s satisfy ψ. The corresponding worlds t in M are those i-accessible from s which satisfy φ. As truth values of formulas may change in an update step, the right description of these worlds in M is not that they satisfy ψ (which they do in $M|\varphi$), but rather $[!\varphi]\psi$: they become ψ after the update. Finally, $!\varphi$ is a partial function (a function that is not everywhere defined), for φ must be true for its announcement to be executable (and if the announcement is not executable, the update result is undefined). Thus, we make our assertion on the right (the assertion about the model after the update) conditional on $!\varphi$ being executable, i.e., on φ being true. Putting this together, $[!\varphi]\Box_i\psi$ says the same as $\varphi \to \Box_i(\varphi \to [!\varphi]\psi)$.

Here is how this functions in a complete calculus of public announcement (we state the theorem without proof):

Theorem 5.67 PAL is axiomatized completely by the laws of epistemic logic over our static model class plus the following recursion axioms:

$$\begin{aligned}
[!\varphi]p &\leftrightarrow \varphi \to p \text{ for atomic facts } p\\
[!\varphi]\neg\psi &\leftrightarrow \varphi \to \neg[!\varphi]\psi\\
[!\varphi](\psi \wedge \chi) &\leftrightarrow [!\varphi]\psi \wedge [!\varphi]\chi\\
[!\varphi]\Box_i\psi &\leftrightarrow \varphi \to \Box_i(\varphi \to [!\varphi]\psi).
\end{aligned}$$

Note that these axioms are all equivalences. To reason with such equivalences we can use the following principle:

Leibniz' principle: If $\vdash \varphi \leftrightarrow \psi$ and χ' is the result of replacing a subformula φ in χ by ψ, then $\vdash \chi \leftrightarrow \chi'$.

This principle is used several times in the following example. One application is the inference from $\vdash [!q]q \leftrightarrow (q \to q)$ to $\vdash (q \to \Box(q \to [!q]q)) \leftrightarrow (q \to \Box(q \to (q \to q)))$.

Example 5.68 (Announcing an atomic fact makes it known) Here is a typical calculation using the axioms (we use $\top$ for a formula that is always true, say $p \vee \neg p$).

$$\begin{aligned}
[!q]\Box q &\leftrightarrow (q \to \Box(q \to [!q]q))\\
&\leftrightarrow (q \to \Box(q \to (q \to q)))\\
&\leftrightarrow (q \to \Box\top)\\
&\leftrightarrow (q \to \top)\\
&\leftrightarrow \top.
\end{aligned}$$

The second step uses the equivalence of $[!q]q$ and $q \to (q \to q)$, the third that of $q \to (q \to q)$ and $\top$, the fourth that of $\Box\top$ and $\top$. To see that $\vdash \Box\top \leftrightarrow \top$, notice that $\vdash \Box\top \to \top$ is an instance of the **T** axiom schema, while from $\vdash \top$ we get $\vdash \Box\top$ by necessitation, and from $\vdash \Box\top$ we get $\vdash \top \to \Box\top$ by propositional logic.

Example 5.69 (Announcement of propositional facts is order-independent)

$$\begin{array}{rcl} [!p][!q]\varphi & \leftrightarrow & [!p](q \to \varphi) \\ & \leftrightarrow & (p \to (q \to \varphi)) \\ & \leftrightarrow & ((p \wedge q) \to \varphi). \end{array}$$

Example 5.70 (Moore sentences again) Let us calculate the conditions under which Moore announcements do make themselves true, using the axioms. First we do a separate calculation to compute $[!(p \wedge \neg\Box p)]\Box p$:

$$\begin{array}{rcl} [!(p \wedge \neg\Box p)]\Box p & \leftrightarrow & (p \wedge \neg\Box p) \to \Box((p \wedge \neg\Box p) \to [!(p \wedge \neg\Box p)]p) \\ & \leftrightarrow & (p \wedge \neg\Box p) \to \Box\top \\ & \leftrightarrow & (p \wedge \neg\Box p) \to \top \\ & \leftrightarrow & \top. \end{array}$$

Next, we are going to use this:

$$\begin{array}{rcl} [!(p \wedge \neg\Box p)](p \wedge \neg\Box p) & \leftrightarrow & [!(p \wedge \neg\Box p)]p \wedge [!(p \wedge \neg\Box p)]\neg\Box p \\ & \leftrightarrow & ((p \wedge \neg\Box p) \to p) \wedge ((p \wedge \neg\Box p) \to \neg[!(p \wedge \neg\Box p)]\Box p) \\ & \leftrightarrow & ((p \wedge \neg\Box p) \to \neg[!(p \wedge \neg\Box p)]\Box p) \\ & \leftrightarrow & ((p \wedge \neg\Box p) \to \bot \\ & \leftrightarrow & \neg p \vee \Box p. \end{array}$$

In the next-to-last line of this derivation we used the fact we proved before: that $[!(p \wedge \neg\Box p)]\Box p \leftrightarrow \top$ is a theorem, and therefore, that $\neg[!(p \wedge \neg\Box p)]\Box p \leftrightarrow \bot$ is a theorem too.

What this derivation says is that update with $!(p \wedge \neg\Box p)$ results in $p \wedge \neg\Box p$ in precisely those cases where the update *cannot be executed* because what it expresses is false.

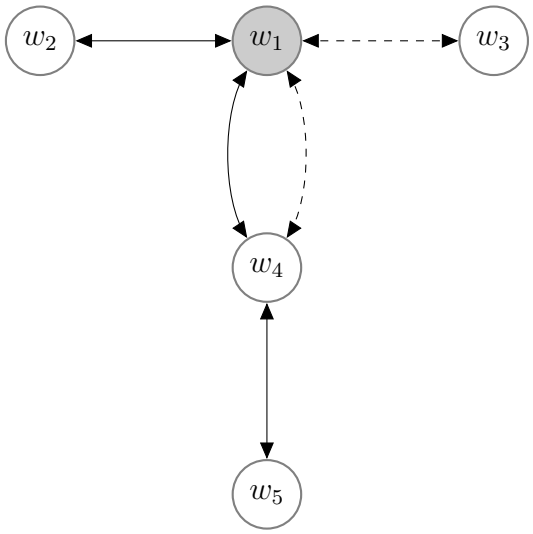
Example 5.71 (Conversation) PAL may be used as a logic of longer-term conversations, or observation procedures, by iterating single update steps. Here is a relevant observation:

Fact 5.72 The formula $[!\varphi][!\psi]\chi \leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\chi$ is valid.

This formula describes how in sequences of two announcements the second announcement is interpreted 'relative' to the update effect of the first.

Optimal communication What can agents in a group achieve by maximal communication? Consider two epistemic agents that find themselves in some collective information state M, at some actual situation s. They can tell each other things they know, thereby cutting down the model to smaller sizes. Suppose they wish to be maximally cooperative.

Example 5.73 (The best agents can do by internal communication) What is the best that can be achieved in the following model? Assume solid links are (symmetric) accessibilities for Q, and dashed links accessibilities for A. Note that in this example the accessibilities are not reflexive.



Geometrical intuition suggests that this must be:

Indeed, a two-step conversation getting here is the following:

- Q sighs: "I don't know".
- Then A sighs: "I don't know either".

It does not matter if you forget details, because it also works in the opposite order.

But maybe we have to assume the accessibilities in the example express belief rather than knowledge, because, as we have seen, knowledge models always have reflexive accessibilities. The accessibilities in the model are not reflexive. If we reinterpret the links in the example model as links expressing belief, the following conversation has the desired effect:

- Q, with indignation: "I don't believe just anything, you know".
- Then A, also indignant: "Well, neither do I".

The first update is with the formula $\neg\Box_Q\bot$, the second with $\neg\Box_A\bot$.

Exercise 5.74 Give equivalent versions for the PAL axioms with existential modalities $\langle!\varphi\rangle$, where $\langle!\varphi\rangle\psi$ is defined as $\neg[!\varphi]\neg\psi$.

A remarkable feature of the axioms for PAL is that the principles about public announcements in the axiomatisation are all equivalences. Also, on the left-hand sides the public announcement operator is the principal operator, but on the righthand sides it is not. What this means is that the axioms reveal that PAL is much more expressive than one might think. It turns out that PAL can encode intricate dynamics of information, provided you take the trouble of analyzing what goes on in information update, in the way we have done.

The principles we have uncovered (in the form of axioms for information update) can be used to 'translate' a formula of PAL to a formula of our standard epistemic language EL. In other words: every statement about the effects of public announcement on individual knowledge is equivalent to a statement about just individual knowledge.

It should be noted, however, that this reduction goes away when we look at temporal processes, protocols and games, the next area one can go from here.

5.10 Outlook — Information, Knowledge, and Belief

From knowledge to belief While information and knowledge are important, our actions are often driven by less demanding attitudes of belief. I ride my bicycle since I believe that it will get me home, even though I can imagine worlds where an earthquake happens. With this distinction in attitude comes one of dynamics. An event of hard information changes irrevocably what I know. If I see the Ace of Spades played on the table, I come to know that no one holds it any more. But there are also events of soft information, which affect my current beliefs without affecting my knowledge in a card game. I see you smile. This makes it more likely that you hold a trump card, but it does not rule out that you do not. How to model all this?

Belief and plausibility models An agent believes what is true, not in all epistemically accessible worlds, but only in the most plausible ones. I believe my bicycle will get me home early, even though I do not know that it will not disappear in an earthquake chasm. But worlds where it stays on the road are more plausible than those where it drops down, and among the former, those where it arrives on time are more plausible than those where it does not.

Definition 5.75 (Epistemic-doxastic models) Epistemic-doxastic models are structures

$$M = (W, \{\sim_i |\ i \in I\}, \{\leq_i |\ i \in I\}, V)$$

where the relations $\sim_i$ stand for epistemic accessibility, and the $\leq_i$ are comparison relations for agents read as follows:

$$x \leq_i y \text{ if agent } i \text{ considers } x \text{ at least as plausible as } y.$$

One can impose several conditions on the plausibility relations, depending on their intuitive reading. An often-used minimum is reflexivity and transitivity, while a lusher version adds

Connectedness For all worlds s, t, either $s \leq t$ or $t \leq s$.

Definition 5.76 (Belief as truth in the most plausible worlds) In epistemic-doxastic models, knowledge is interpreted as usual, while we now say that

$$M, s \models B_i \varphi$$

iff $M, t \models \varphi$ for all worlds t that are minimal in the ordering $\leq_i$.

This can be further refined, as follows.

Definition 5.77 (Conditional Belief as Plausibility Conditionals) Extend the language with conditional belief formulas $B_i^{\psi}\varphi$, with the intuitive reading that, conditional on ψ, the agent believes that φ. Formally:

$$M, s \models B_i^{\psi}\varphi \quad \text{iff} \quad M, t \models \varphi \text{ for all worlds } t \text{ which are minimal for the ordering } \leq_i \text{ in the set } \{u \mid M, u \models \psi\}.$$

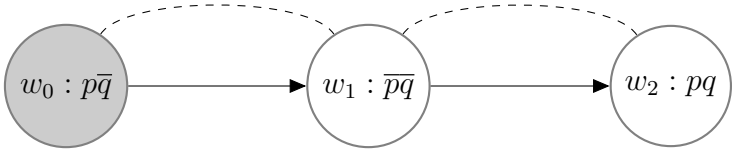
Belief change under hard information The capacity for learning from new facts contradicting our earlier beliefs seems typical of rational agency.

Fact 5.78 The formula

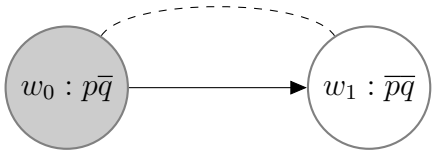
$$[!\varphi]B^{\psi}\chi \leftrightarrow (\varphi \rightarrow B^{\psi}[!\varphi]\chi)$$

is valid for beliefs after hard information.

Example 5.79 (Misleading with the truth) Consider a model where an agent believes that p, which is indeed true in the actual world to the far left, but for 'the wrong reason', viz. she thinks the most plausible world is the one to the far right. For convenience, assume that the final world also verifies a unique proposition letter q. The dashed links are knowledge links, the solid arrows are plausibility arrows, for the same agent.



Now giving the true information that we are not in the final world ($\neg q$) updates to:



in which the agent believes mistakenly that $\neg p$.

5.11 Outlook – Social Knowledge

Example 5.80 Imagine two generals who are planning a coordinated attack on a city. The generals are on two hills on opposite sides of the city, each with their armies, and they know they can only succeed in capturing the city if the two armies attack at the same time. But the valley that separates the two hills is in enemy hands, and any messenger that is sent from one army base to the other runs a severe risk of getting captured. The generals have agreed on a joint attack, but they still have to settle the time.

So the generals start sending messengers. General 1 sends a soldier with the message "We will attack tomorrow at dawn". Call this message p. Suppose his messenger gets across to general 2 at the other side of the valley. Then $\Box_2 p$ holds, but general 1 does not know this because he is uncertain about the transfer of his message. Now general 2 sends a messenger back to assure 1 that he has received his message. Suppose this messenger also gets across without being captured, then $\Box_1\Box_2 p$ holds. But general 2 does not know this, for he is uncertain about the success of transfer: $\neg\Box_2\Box_1\Box_2 p$. General 1 now sends a second messenger. If this one also safely delivers his message we have $\Box_2\Box_1\Box_2 p$. But general 1 does not know this, and so on, and so on. In this way, they'll continue sending messages infinitely (and certainly not attack tomorrow at dawn).

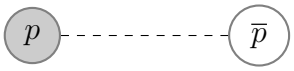
Clearly, this procedure will never establish common knowledge between the two generals. They share the knowledge of p but that is surely not enough for them to be convinced that

they will both attack at dawn. In case of real common knowledge every formula of the infinite set

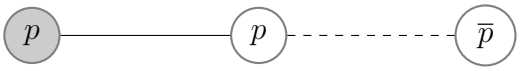
$$\{\Box_1 p, \Box_2 p, \Box_1\Box_2 p, \Box_2\Box_1 p, \Box_1\Box_2\Box_1 p, \Box_2\Box_1\Box_2 p, \ldots\}$$

holds.

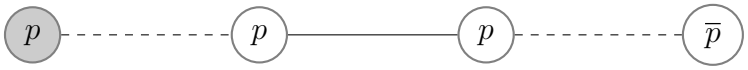
Here are pictures of how the situation as given in the previous example develops after each messenger delivers his message. Initially, general 1 settles the time of the attack. He knows that p but he also knows that general 2 does not know (with a dashed link for the accessibility of general 2):



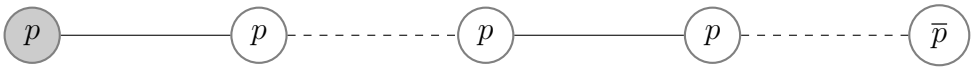
After the first messenger from 1 to 2 gets safely across we have (with a solid link for the accessibility relation of general 1):



After the message of 2 to 1 is safely delivered we get:



Successful transfer of the second message from 1 to 2 results in:



Note that in the second world from the left it does not hold that $\Box_2\Box_1\Box_2 p$, and therefore $\neg\Box_1\Box_2\Box_1\Box_2 p$ is true in the actual world.

The example makes it seem that achieving common knowledge is an extremely complicated or even impossible task. This conclusion is too negative, for common knowledge can be established immediately by public announcement. Suppose the two generals take a risk and get together for a meeting. Then general 1 simply says to general 2 "We will attack tomorrow at dawn", and immediately we get:

Still, we cannot express common knowledge between 1 and 2 by means of a single formula of our language. What we want to say is that the stacking of knowledge operators goes on indefinitely, but we have no formula for this.

The way to handle this is by adding a modality of common knowledge. $C_G\varphi$ expresses that it is common knowledge among the members of group G that φ. Here is the truth definition for it:

$$M, s \models C_G\varphi \quad \text{iff} \quad \begin{array}{l}\text{for all } t \text{ that are reachable from } s \\ \text{by some finite sequence of } \rightarrow_i \text{ steps } (i \in G)\text{: } M, t \models \varphi.\end{array}$$

Theorem 5.81 The complete epistemic logic with common knowledge is axiomatized by adding two axioms and a rule to the minimal epistemic logic. In the two axioms, E_G is used as an abbreviation for everybody in the group knows (defined as $E_G\varphi \leftrightarrow \Box_{g_1}\varphi \wedge \cdots \wedge \Box_{g_n}\varphi$, for all $g_1, \ldots, g_n$ in G):

Fixed-Point Axiom $C_G\varphi \leftrightarrow (\varphi \wedge E_G C_G\varphi)$.

Induction Axiom $(\varphi \wedge C_G(\varphi \rightarrow E_G\varphi)) \rightarrow C_G\varphi$.

C Necessitation Rule If φ is a theorem, then $C_G\varphi$ is also a theorem.

The axioms are also of independent interest for what they say. The Fixed-Point Axiom expresses an intuition of reflective equilibrium: common knowledge of φ is a proposition X implying φ of which every group member knows that X is true. On top of this, the Induction Axiom says that it is not just any equilibrium state of this kind, but the largest one.

To axiomatize PAL with common knowledge we need more expressive power. One possible (and elegant) way to achieve this is by adding an operator for *conditional common knowledge*, $C_G^\varphi\psi$, with the following truth definition:

$$M, s \models C_G^\varphi\psi \quad \text{iff} \quad \begin{array}{l}\text{for all } t \text{ that are reachable from } s \\ \text{by some finite sequence of } \rightarrow_i \text{ steps } (i \in G) \\ \text{through a series of states that all satisfy } \varphi \\ \text{it holds that } M, t \models \psi.\end{array}$$

This allows for a complete axiomatisation (again, we state the theorem without proof):

Theorem 5.82 PAL with conditional common knowledge is axiomatized completely by adding the valid reduction law

$$[!\varphi]C_G^\psi\chi \leftrightarrow (\varphi \rightarrow C_G^{\varphi \wedge [!\varphi]\psi}[!\varphi]\chi).$$

Example 5.83 Many social rituals are designed to create common knowledge. A prime example is cash withdrawal from a bank. You withdraw a large amount of money from your bank account and have it paid out to you in cash by the cashier. Typically, what happens is this. The cashier looks at you earnestly to make sure she has your full attention, and then she slowly counts out the banknotes for you: one thousand (counting ten notes while saying *one*, *two*, *three*, ..., *ten*), two thousand (counting another ten notes), three thousand (ten notes again), and four thousand (another ten notes). This ritual creates common knowledge that forty banknotes of 100 euros were paid out to you.

To see that this is different from mere knowledge, consider the alternative where the cashier counts out the money out of sight, puts it in an envelope, and hands it over to you. At home you open the envelope and count the money. Then the cashier and you have knowledge about the amount of money that is in the envelope. But the amount of money is not common knowledge among you. In order to create common knowledge you will have to insist on counting the money while the cashier is looking on, making sure that you have her full attention. For suppose you fail to do that. On recounting the money at home you discover there has been a mistake. One banknote is missing. Then the situation is as follows: the cashier believed that she knew there were forty banknotes. You now know there are only thirty-nine. How are you going to convince your bank that a mistake has been made, and that it is their mistake?

5.12 Outlook – Secrecy and Security

In computer science, protocols are designed and studied that do not reveal secret information to eavesdroppers. A strong property of such protocols is the following:

> Even if all communication is overheard, the secret is not compromised.

One example of how this can be achieved is given by the so-called *Dining Cryptographers Protocol*, designed by computer scientist David Chaum. The setting of this protocol is a situation where three cryptographers are eating out. At the end of the dinner, they are informed that the bill has been paid, either by one of them, or by NSA (the National Security Agency). Respecting each others' rights to privacy, they want to find out whether NSA paid or not, in such a way that in case one of them has paid the bill, the identity of the one who paid is not revealed to the two others.

They decide on the following protocol. Each cryptographer tosses a coin with his right-hand neighbour, with the result of the toss remaining hidden from the third person. Each cryptographer then has a choice between two public announcements: that the coins that she has observed agree or that they disagree. If she has not paid the bill she will say that they agree if the coins are the same and that they disagree otherwise; if she has paid the bill she will say the opposite: she will say that they agree if in fact they are different and she will say that they disagree if in fact they are the same. If everyone is speaking the

truth, the number of 'disagree' announcements will be even. This reveals that NSA has picked up the bill. If one person is 'lying', the number of 'disagree' announcements will be odd, indicating that one among them has paid the bill.

One can analyse this with epistemic logic by starting out with a model where the diners have common knowledge of the fact that either NSA or one of them has paid. Next, one updates with the result of the coin tosses, and with communicative acts representing the sharing of information between a cryptographer and his neighbour about these results.

Calling the cryptographers 1, 2 and 3, use p_1, p_2 and p_3 to express that 1, 2 or 3 has paid. The aim of the protocol is that everybody learns whether the formula $p_1 \vee p_2 \vee p_3$ is true or not, but if the formula is true, nobody (except the payer herself) should learn which of the three propositions was true. It is left to you to figure out why the above protocol achieves this goal.

Summary of Things You Have Learnt in This Chapter *You have learnt to look at information as uncertainty between various possible states of affairs, for cases of a single agent, but also for multi-agent settings that involve knowledge about knowledge. You know what information models are, and you are able to evaluate formulas from epistemic logic in information models. You have some experience with constructing formal proofs in epistemic logic. You are familiar with the concept of information update, and you can understand simple protocols designed to update information states. You have grasped the distinction between individual knowledge and common knowledge, and know in which cases public announcements can be used to establish common knowledge.*

Further Reading A classic on the logic of knowledge and belief is Jaakko Hintikka's [Hin62]. Epistemic logic for computer science is the subject of [MvdH95] and [FHMV95]. A textbook treatment of dynamic epistemic logic can be found in [DvdHK06]. A recent book on information exchange and interaction is [vB11].