5 Epistemic Logic

5.1 Epistemic possibility

When we say that something is possible, we often mean that it is compatible with our information. This "epistemic" flavour of possibility is studied in epistemic logic. More generally, epistemic logic provides tools to formally reason about knowledge, belief, information, communication, and related concepts. Originating in philosophy in the 1950s and 1960s, this branch of modal logic has found many applications in game theory, computer science, cognitive science, and other disciplines.

Standard epistemic logic heavily relies on the possible-worlds semantics introduced in chapters 2 and 3. The guiding intuition is that information rules out possibilities. The more you know, the fewer possibilities are left open by your knowledge. For an omniscient agent, only one world is epistemically possible (compatible with her knowledge): the actual world. For an agent who knows nothing at all, every world is epistemically possible.

Knowledge and information vary not just from world to world, but also from person to person. The detective doesn't know who stole the jewels, but the thieve does. If we want to reason about the information available to different agents, we must keep track of who knows what. A world that's accessible (i.e., epistemically possible) for the detective need not be accessible for the thief. So we often need multiple accessibility relations, one for each agent.

Definition 5.1

A **multi-modal Kripke model** consists of

- a non-empty set W,
- a finite set of binary relation $R_1, R_2, \ldots, R_n$ on W, and
- a function V that assigns to each sentence letter and each element of W a truth-value.

Each accessibility relation R_i represents the information available to a particular agent. A world v is R_1-accessible from w iff v is compatible with the information agent 1 has at world w.

If we have multiple accessibility relations, we also need multiple boxes and diamonds. So we will expand the language $\mathfrak{L}_M$ by introducing several operators $\Box_1, \Box_2, \Box_3, \ldots$ with their duals $\Diamond_1, \Diamond_2, \Diamond_3, \ldots$. We can then say things like

$$\Diamond_1 p \wedge \neg \Diamond_2 p,$$

meaning that some p-world is accessible for agent 1 while no p-world is accessible for agent 2. How many accessibility relations and operators we need depends on the kinds of situations we want to model.

The definition of truth at a world in a Kripke model (definition 3.2) is easily extended to multi-modal Kripke models. Instead of clauses (g) and (h), we have the following conditions, for each pair of a modal operator $\Box_i/\Diamond_i$ and the corresponding accessibility relation R_i:

$M, w \models \Box_i A$ iff $M, v \models A$ for all $v \in W$ such that wR_iv.
$M, w \models \Diamond_i A$ iff $M, v \models A$ for some $v \in W$ such that wR_iv.

To remind ourselves that we are dealing with epistemic modality, it is customary in epistemic logic to write $\Box_1, \Box_2$, etc. as 'K_1','K_2', etc. (or 'K_a','K_b', etc.). For once, the letter 'K' here stands not for Kripke but for knowledge. There is no established convention for the duals $\Diamond_i$. I will use 'M_i'; others use 'P_i', '$\langle \mathsf{K}_i \rangle$', or '$\langle i \rangle$'. So to express that the information available to agent 2 is incompatible with p, I would write '$\neg\, \mathsf{M}_2\, p$' or '$\mathsf{K}_2\, \neg p$'.

Informally, 'K_i' may be read as 'agent i knows that', and 'M_i' as 'for all agent i knows, it might be that'. However, this translation must be taken with a grain of salt, as the K operators of standard epistemic logic formalize a conception of knowledge that does not perfectly match the use of 'knowledge' in ordinary language.

To see why, note that if some propositions are true at a world, then anything that logically follows from these propositions is also true at that world (by definition 3.2). For example, if p and q are true at w, then so is $p \wedge q$. As a consequence, if p and q are true at all R_i-accessible worlds, then $p \wedge q$ is also true at all these worlds. So if $\mathsf{K}_i\, p$ and $\mathsf{K}_i\, q$ are both true, then Kripke semantics guarantees that $\mathsf{K}_i(p \wedge q)$ is true. More generally, the knowledge operators in Kripke semantics are **closed**

under logical consequence, meaning that if B logically follows from $A_1, \ldots, A_n$, and $\mathsf{K}_i\, A_1, \ldots, \mathsf{K}_i\, A_n$, then $\mathsf{K}_i\, B$.

Our ordinary conception of knowledge is not closed under logical consequence. If you know the axioms of a mathematical theory, you don't automatically know everything that logically follows from these axioms. On the other hand, suppose Alice knows that Bob is in the living room. Intuitively, we can infer that Alice knows that *someone* is in the living room – even if she never consciously entertained that conclusion. Our everyday conception of knowledge seems to be closed under "obvious logical consequence": if an agent knows $A_1, \ldots, A_n$, and B is an obvious logical consequence of $A_1, \ldots, A_n$, then the agent knows B. But that can't be right, for closure under obvious logical consequence entails closure under non-obvious logical consequence.

Exercise 5.1

Let's say that B is an *obvious logical consequence* of $A_1, \ldots, A_n$ iff B can be inferred from $A_1, \ldots, A_n$ by one application of Modus Ponens. Suppose an agent knows all instances of all axioms of propositional logic. If knowledge is closed under obvious logical consequence, then the agent knows everything that logically follows from these axioms. Explain why.

We won't try to figure out the exact rules that govern our ordinary conception of knowledge. Let's say that an agent *implicitly knows* a proposition if the proposition logically follows from things the agent knows. Implicit knowledge is, by definition, closed under logical consequence. And it is a useful concept. Roughly speaking, it captures what information an agent has about the world. If what you know logically entails p, then the information you have about the world settles that p is the case. When we speak of knowledge in epistemic logic, we normally mean implicit knowledge.

Exercise 5.2

Translate the following sentences into the language of epistemic logic, ignoring my warnings about the mismatch between K and the ordinary concept of knowledge.

(a) Alice knows that it is either raining or snowing.

(b) Either Alice knows that it is raining or that it is snowing.
(c) Bob knows whether it is raining.
(d) Carol knows that she doesn't know that it is raining.
(e) Alice knows that Bob knows whether it is raining.

Exercise 5.3

The ordinary concept of knowledge logically ill-behaved in more than one respect. Let K^* be an operator that applies to a sentence A iff we would intuitively say that an agent knows A. Assume the agent in question knows the axioms of ZFC set theory. Define K^+ as the logical closure of K^*; that is,

$\mathsf{K}^+ A \Leftrightarrow_{\text{def}}$ A is entailed by sentences $A_1, \ldots, A_n$ such that $\mathsf{K}^* A_1, \ldots, \mathsf{K}^* A_n$.

Note the similarity between K^+ and the mathematical provability operator from section 4.4. Indeed, with minimal further assumptions one can prove that K^+ validates the **GL** schema:

$$\mathsf{K}^+(\mathsf{K}^+ A \rightarrow A) \rightarrow \mathsf{K}^+ A$$

From the definition of K^+, it follows that

$$\mathsf{K}^*(\mathsf{K}^+ A \rightarrow A) \rightarrow \mathsf{K}^+ A.$$

Explain why this is an intuitively unacceptable principle about knowledge.

5.2 Gaining information

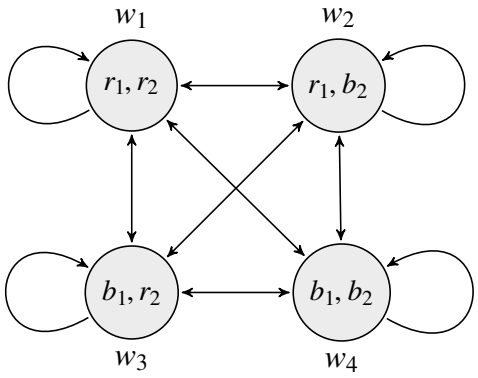
To get a feeling for epistemic logic, it helps to work through a few examples in which agents gain information. Let's start with a simple case, with only one agent; so we can use standard Kripke models with a single accessibility relation.

Two cards are drawn from a deck of red and black cards and put face-down in front of Ava. There are four possibilities:

- Both cards are red.

- Both are black.
- Card 1 is red and card 2 black.
- Card 1 is black and card 2 red.

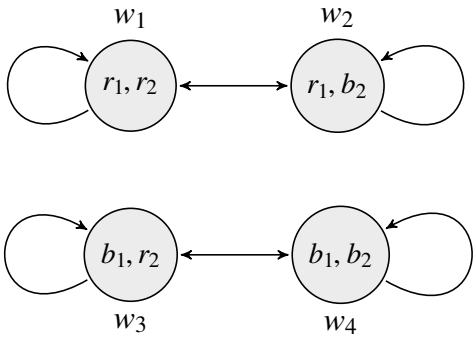
Since we are only interested in Ava's attitude towards the colour of her cards, we can take these possibilities as our possible worlds. (It usually suffices to make the worlds in our models maximally specific *with respect to the questions we're interested in*.) Here is the situation pictured as a Kripke model:



Every world has access to every other world as well as to itself. For example, in world w_1 both cards are red, but Ava doesn't know this. Ava's information state in world w_1 is compatible with all four possibilities.

Note the division of labour between V and R. The interpretation function V fixes the truth-value of the non-modal sentences at each world. In the present example, V tells us the colour of each card at each world. It says nothing about Ava's knowledge. Information about what is known at the various worlds is represented by the accessibility relation.

Now assume Ava turns over the first card. Without knowing what she sees, we can say how the model changes.

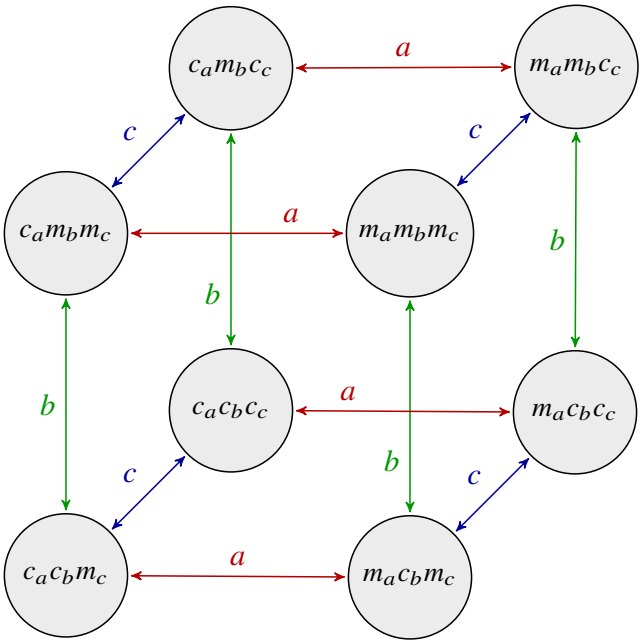


In w_1, Ava sees that the first card is red. So she can rule out worlds w_3 and w_4, where the first card is black. That's why there are no more arrows from w_1 to these worlds. But she still can't rule out world w_2, where the second card is black. If Ava then turns over the second card, each world becomes accessible only from itself: at each world, Ava knows the colour of her cards.

Let's look at a more interesting case, known as the *Muddy Children* puzzle.

> Three (intelligent) children have been playing outside. They can't see or feel if their own face is muddy, but they can see who of the others has mud on their face. Coming inside, mother tells them: "at least one of you has mud on their face". She then asks, "do you know if you have mud on your face?". All three children say no. Mother asks again, "do you know if you have mud on your face?". This time, two children say, yes; one says no. What happens when the mother asks the question a third time? And how many children have mud on their face?

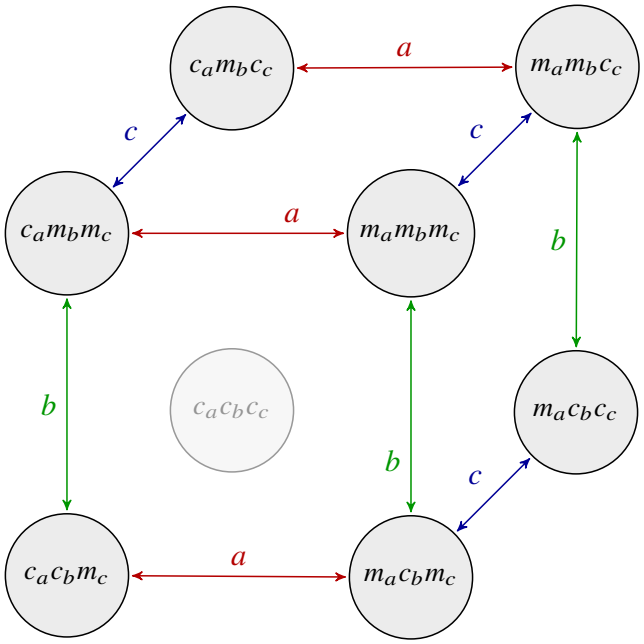
Let's draw a model. I'll call the three children Alice, Bob, and Carol, and I'll use m_a, m_b, m_c as sentence letters expressing, respectively, that Alice/Bob/Carol is muddy; c_a, c_b, c_c mean that Alice/Bob/Carol is clean. Before the mother's first announcement, there are eight possibilities.



Since we have three epistemic agents, we have three accessibility relations. I have left out the (three) arrows from each world to itself, to remove clutter.

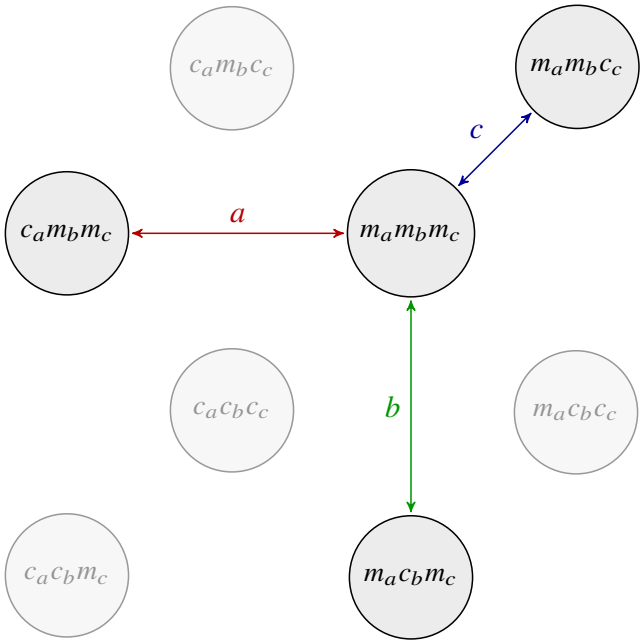
The model reflects the fact that each child can see the others. For example, at the top left world ($c_a m_b c_c$), Alice sees that Bob is muddy while Carol is clean; consequently, the only epistemic possibilities for Alice at that world are the two worlds at the top: $c_a m_b c_c$ itself and $m_a m_b c_c$. In general, the only accessible worlds for a given child at a given world w are worlds at which the other children's state of muddiness is the same as at w.

What changes through the mother's first announcement, "at least one child has mud on their face"? The announcement tells *us* that we're not in the $c_a c_b c_c$ world. More importantly, it allows *each child* to rule out the $c_a c_b c_c$ world (since they all hear and accept the announcement).



Next, the mother asks if anyone knows whether they are muddy. No child says yes. So no-one knows whether they are muddy. And everyone now knows that no-one knows whether they are muddy. We can go through the above seven possibilities to see if at any of them, anyone knows whether they are muddy. At the top left world ($c_a m_b c_c$) Alice doesn't know whether she is muddy, because the $m_a m_b c_c$ world (top right) is a-accessible; nor does Carol know whether she is muddy, because $c_a m_b m_c$ is c-accessible. But Bob knows that he is muddy: no other world is b-accessible. Intuitively, at the $c_a m_b c_c$ world, Bob sees two clean children (Alice and Carol), and he has just been told that not all children are clean. So he can infer that he is muddy. But we know that Bob didn't say that he knows whether he is muddy. So we (and all the children) can rule out the top left world as an open possibility.

By the same reasoning, every world connected with only two arrows to other worlds can be eliminated at this stage.



When the mother asks again if anyone knows whether they are muddy, two children say "yes". So everyone comes to know that two children know whether they are muddy. In the middle world of the above model ($m_a m_b m_c$), however, no child knows whether they are muddy. So that world is not actual, and it is no longer accessible for anyone. The remaining open possibilities are $c_a m_b m_c$, $m_a c_b m_c$, and $m_a m_b c_c$, each of which is only accessible from itself.

Now we can answer the questions. In the three remaining worlds, every child now knows who is muddy and who is clean. So if the mother asks her question for the third time, everyone says yes. Also, exactly two children have mud on their face.

Exercise 5.4

Albert and Bernard just met Cheryl. "When is your birthday?", Albert asks. Cheryl answers, "I'll give you some clues". She writes down a list of 10 dates:

15 May, 16 May, 19 May
17 June, 18 June
14 July, 16 July
14 August, 15 August, 17 August

"My birthday is one of these", she says. Then she announces that she will whisper the month of her birthday in Albert's ear and the day in Bernard's. After the whispering, she asks Albert if he knows her birthday. Albert says, "no, but I know that Bernard doesn't know either". To which Bernard responds: "Right. I didn't know until now, but now I know". Albert: "Now I know too!" Draw a (multi-modal) Kripke model for each stage of the conversation. When is Cheryl's birthday?

5.3 The logic of knowledge

What is the logic of (implicit) knowledge? That is, which sentences in the language of epistemic logic should count as logically valid, and which sentences should we treat as logical consequences of which other sentences?

One obvious assumption is that knowledge implies truth. If you know that it is raining, we can infer that it is raining. So the **T**-schema should be valid:

(**T**) $\quad \mathsf{K}_i\, A \to A$

In chapter 3, we saw that the **T**-schema is valid on all and only the reflexive frames. We will therefore assume that every accessibility relation R_i in every epistemic Kripke model is reflexive. Non-reflexive Kripke models are unsuitable as models for knowledge.

Reflexivity implies seriality, which corresponds to the schema

(**D**) $\quad \mathsf{K}_i\, A \to \mathsf{M}_i\, A$

Intuitively, this means that the information available to an agent is never contradictory: if the information entails A (as $\mathsf{K}_i\, A$ asserts), then it does not entail $\neg A$ (i.e, $\neg\, \mathsf{K}_i\, \neg A$).

Let's look at some other schemas from earlier chapters. I will omit the subscript 'i' in what follows, since these schemas involve only one kind of box and diamond.

One noteworthy schema is **4**, which corresponds to transitivity of the accessibility relation.

(**4**) $\quad \mathsf{K}\, A \to \mathsf{K}\,\mathsf{K}\, A.$

In epistemic logic, **4** is known as **positive introspection** or **the KK principle**. If you (implicitly) know something, does it follow that you (implicitly) know that you (implicitly) know it? Many arguments have been given for either side.

A well-known argument against the KK principle is based on the idea that knowledge requires "safety": you know p only if you couldn't easily have been wrong about p. The safety condition can be motivated by Gettier cases. Suppose you are looking at the only real barn in a valley which, unbeknownst to you, is full of fake barns. Your belief that you're looking at a barn is true, and it seems to be justified. But intuitively, it isn't knowledge. You don't know that what you're looking at is a real barn. Why not? Advocates of the safety condition suggest that you don't have knowledge because you could easily have been wrong. You truly know p only if there is no "nearby" possibility at which p is false, where "nearness" is a matter of similarity in certain respects.

On the safety account, you know that you know p only if there is no nearby world at which you don't know p. That is, you know at world w that you know p only if you know p at all worlds v that are relevantly similar to w. And you know p at v only if p is true at all worlds u that are relevantly similar to v. But similarity isn't transitive: the fact that u is similar to v and v is similar to w does not entail that u is similar to w. So it can happen that p holds at all "nearby" worlds, but not at all worlds that are nearby from a nearby world. In that case, you may know p without knowing that you know p.

Not everyone accepts the safety condition. Other accounts of knowledge vindicate the KK principle. For example, some have argued that an agent knows p (roughly) iff the agent is in a belief state which *indicates* p, in the sense that

(1) under normal conditions, being in that state implies p, and
(2) conditions are normal.

We can formalize this concept in modal logic. Let N mean that conditions are normal (whatever exactly that means), and let $\Box$ be an operator that quantifies unrestrictedly over all possible worlds. $\Box(N \to A)$ then means that A is true at all world at which conditions are normal. According to the above definition, a state s indicates p iff

(*) $\qquad \Box(N \to (s \to p)) \wedge N.$

So s indicates that s indicates that p iff

$$(**) \qquad \Box(N \to (s \to (\Box(N \to (s \to p)) \wedge N))) \wedge N.$$

A quick tree proof reveals that (*) entails (**). That is, whenever a state indicates p, it also indicates that it indicates p. On the indication account of knowledge, a belief state that constitutes knowledge therefore automatically constitutes knowledge of knowledge: the **4** schema is valid.

Exercise 5.5

Give an S5 tree proof to show that (*) entails (**). Why can we assume S5 here?

The KK principle says that people always know what they know. We might similarly postulate that people always know what they *don't* know. This would give us schema **5**, or **negative introspection**.

$$(\mathbf{5}) \qquad \neg \mathsf{K}\, A \to \mathsf{K} \neg \mathsf{K}\, A.$$

Semantically, the **5**-schema corresponds to euclidity. Since reflexivity and euclidity entail symmetry (exercise 3.3), positive and negative introspection, together with the assumption that knowledge entails truth, would make the accessibility relation an equivalence relation; the logic of knowledge would be S5.

S5 is the simplest of the (non-trivial) modal logics, which may be one reason why negative and positive introspection are widely assumed in theoretical computer science. The assumptions can also be justified by certain ideas about the systems we are trying to model. Imagine an artificial agent whose database can store a finite number of propositions $p_1, \ldots, p_n$. The agent receives information through a reliable channel, so that the agent is guaranteed to never store any false information. We might then say that the agent *knows* A just in case A is entailed by what the agent has stored in its database. By scanning its own database, the agent can easily find out whether or not it knows p_i. So if the agent knows something, then it is in a position to know that it knows it (**4**). Similarly, if the agent doesn't know something, then it is in a position to know that it doesn't know it (**5**).

In philosophy, however, negative introspection is almost universally rejected.

As Donald Rumsfeld pointed out, there are not only "known unknowns" but also "unknown unknowns". An unknown unknown is something we don't know of which we don't know that we don't know it – precisely the kind of state ruled out by negative introspection.

One way unknown unknowns can come about is through false beliefs. Suppose you believe that p, but p is false. Then you don't know that p, because knowledge implies truth. But will you always know that you don't know that p? Clearly not. On the contrary, if you falsely believe that p, you are likely to falsely believe that you know that p. So negative introspection seems to rule out the possibility of false belief.

Here it's important to not be misled by another curiosity of ordinary language: when we say that someone doesn't know p, this seems to imply that p is true. For example, if I told you that my neighbour doesn't know that I have a pet aardvark, you could reasonably infer that I have a pet aardvark. But it is not clear what licenses this inference. After all, one can only know what is true. So if I *don't* have a pet aardvark then certainly my neighbour doesn't know that I have one. Accordingly, in epistemic logic, $\neg \mathsf{K}\, A$ does not imply A.

Let's say that an agent is *ignorant of* p if they don't know that p and p is true. In ordinary language, saying that someone doesn't know p conveys that the agent in question is ignorant of p. What Rumsfeld had in mind when he spoke of unknown unknowns aren't just cases in which we don't know that we don't know something, but cases in which we are ignorant of our own ignorance.

Exercise 5.6

Rumsfeld said that there are "known unknowns" and "unknown unknowns". But if an "unknown" is something of which we're ignorant, then arguably there are only unknown unknowns. Prove that if the logic of knowledge is at least as strong as K, then ignorance of A entails ignorance of ignorance of A.

Let's return to the logic of knowledge. So far, we have seen that **T** is valid, **4** is controversial, and **5** is plausibly invalid. A common view among epistemic logicians is that the logic of (implicit) knowledge lies somewhere in between S4 and S5, because it validates **T** and **4**, does not validate **5**, but does validate some further principles that are contained in S5 and not in S4.

One way to motivate these further principles is to recall why we rejected negative introspection. If an agent falsely believes p, then they don't know p, and yet they normally don't know that they don't know p. Now, if p is true, then obviously one can't *falsely* believe p. So one might suggest that if p is true and an agent doesn't know p, then they always know that they don't know p. This would give us a schema known as 0.4:

(**0.4**) $\quad (A \wedge \neg \mathsf{K}\, A) \rightarrow \mathsf{K} \neg \mathsf{K}\, A$

The **0.4** schema is S5-valid, but not S4-valid. Adding it to S4 leads to a system known as S4.4.

A more modest extension of S4 adds principle **G**:

(**G**) $\quad \mathsf{M}\mathsf{K}\, A \rightarrow \mathsf{K}\mathsf{M}\, A$

The resulting logic is called S4.2; it is weaker than S4.4 but stronger than S4. We will meet an argument in favour of **G** in the next section.

Exercise 5.7

Use the tree method to confirm the following statements.

(a) $\models_{S4} \mathsf{M}\, A \leftrightarrow \mathsf{M}\mathsf{M}\, A$.
(b) $\models_{S4} \mathsf{M}\mathsf{K}\mathsf{M}\, A \rightarrow \mathsf{M}\, A$.
(c) $\models_{S4} \mathsf{M}\mathsf{K}(A \rightarrow \mathsf{K}\mathsf{M}\, A)$.
(d) $\models_{B} \mathsf{M}\mathsf{K}\, A \rightarrow \mathsf{K}\mathsf{M}\, A$.
(e) $\models_{S4.2} \mathsf{M}\mathsf{K}\, A \wedge \mathsf{M}\mathsf{K}\, B \rightarrow \mathsf{M}\mathsf{K}(A \wedge B)$.

Exercise 5.8

Give an S4 tree proof to show that $(A \wedge \neg \mathsf{K}\, A) \rightarrow \mathsf{K} \neg \mathsf{K}\, A$ and $(\neg A \wedge \mathsf{M}\, A) \rightarrow \mathsf{K}\mathsf{M}\, A$ (both of which are covered by 0.4) together entail **G**.

Exercise 5.9

Explain why Gettier cases cast doubt on the validity of **0.4**.

What if we have several agents, and thus several knowledge operators $\mathsf{K}_1, \mathsf{K}_2$, etc.? Individually, each of these operators should satisfy whatever conditions we want to impose on the logic of knowledge. But are there also new principles governing the interaction between different knowledge operators?

For example, we plausibly want the following to come out valid:

$$\mathsf{K}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{K}_1\,A.$$

If I know that you know that it's raining, then I also know that it's raining. Principles like this, containing different modal operators (that are not definable in terms of each other), are called **interaction principles**.

The standard assumption in epistemic logic is that there are no new interaction principles for the knowledge of several agents – no principles that don't already follow from the logic of individual knowledge. For the above example, it is easy to see that if the **T** schema is valid for K_2, then $\mathsf{K}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{K}_1\,A$ comes out valid, so we don't need to add a separate principle. Think of the relevant Kripke models. Suppose, as $\mathsf{K}_1\,\mathsf{K}_2\,A$ asserts, that A holds at each world that is R_2-accessible from any R_1-accessible world. If the **T** schema is valid for K_2, then any R_1-accessible world is R_2-accessible from itself. It follows that A holds at each R_1-accessible world. So $\mathsf{K}_1\,A$ is true.

Exercise 5.10

Which of the following interaction principles are valid if the logic of individual knowledge is S4?

(a) $\mathsf{M}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{M}_1\,A$
(b) $\mathsf{M}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{M}_2\,\mathsf{M}_1\,A$
(c) $\mathsf{M}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{M}_2\,\mathsf{K}_1\,A$
(d) $\mathsf{K}_1\,\mathsf{K}_2\,A \rightarrow \mathsf{K}_2\,\mathsf{K}_1\,A$

We can also define some new modalities for groups of agents. Let's say that a proposition is **mutually known** in a group G iff it is known by every member of the group. Let E_G be an operator for mutual knowledge. Clearly, $\mathsf{E}_G\,A$ can be defined as $\mathsf{K}_1\,A \wedge \mathsf{K}_2\,A \wedge \ldots \wedge \mathsf{K}_n\,A$, where $\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_n$ are the knowledge operators for the members of the group. So we can't say anything new with the help of E_G. But it can

be instructive to see how E_G behaves depending on the behaviour of the underlying operators $\mathsf{K}_1, \mathsf{K}_2$, etc. For example, if each individual knowledge operator validates the **T** schema, then so does E_G; but if each K_i validates **4** (positive introspection), it does not follow that E_G validates **4**. As a counterexample, consider a group of two agents; both know p, and both satisfy positive introspection, but agent 1 does not know that agent 2 knows p. Then $\mathsf{E}_G\, p$ but $\neg\, \mathsf{E}_G\, \mathsf{E}_G\, p$.

Exercise 5.11

Give an example to show that if each K_i validates **5**, it does not follow that E_G validates **5**.

A more interesting concept that is widely used in many areas of research is that of **common knowledge**. A proposition is commonly known in a group if everyone knows it, everyone knows that everyone knows it, everyone knows that everyone knows that everyone knows it, and so on forever. Let's use C_G as an operator for common knowledge. C_G is not definable in terms of $\mathsf{K}_1, \ldots, \mathsf{K}_n$. Semantically, $\mathsf{C}_G\, A$ is true at a world w iff A is true at all worlds that are reachable from w by some finite sequence of steps following R_1, R_2, ..., or R_n.

By definition, common knowledge validates **4**. It validates **T** whenever individual knowledge validates **T**. So the logic of common knowledge is at least S4. The complete logic of individual and common knowledge turns out to also contain the following (non-trivial) interaction principles, which are easiest to state in terms of E_G:

(**CK1**) $\quad \mathsf{C}_G\, A \leftrightarrow (A \wedge \mathsf{E}_G\, \mathsf{C}_G\, A)$

(**CK2**) $\quad (A \wedge \mathsf{C}_G(A \rightarrow \mathsf{E}_G)) \rightarrow \mathsf{C}_G\, A$

You may want to confirm that these are sound. (They also provide a complete axiomatization of common knowledge when added to an axiomatic calculus for individual knowledge, but that is much harder to see.)

5.4 Knowledge, belief, and other modalities

Issues in the logic of knowledge can sometimes be clarified by looking at the connections between knowledge and belief. To formalise these connections, let's introduce a new operator B for belief – or rather, for *implicit belief*, since B, like K, will be closed under logical consequence. If we wanted to reason about several agents, we would have multiple B operators $\mathsf{B}_1, \mathsf{B}_2, \ldots$, but in this section I am going to focus on a single agent. So we will work with a bi-modal language with two box-like operators, written 'B' and 'K'.

The logic of B is different from the logic of K, if only because beliefs can be false. As a consequence, the **T**-schema is not valid for B. We may, however, accept the weaker **D**-schema,

(**D**) $\quad \mathsf{B}\,A \rightarrow \neg\,\mathsf{B}\,\neg A$

This would mean that if one believes a proposition A then one can't also believe its negation $\neg A$.

In the previous section, I gave some arguments suggesting that knowledge does not validate **4** and **5**. None of these arguments carry over to belief. Many epistemic logicians therefore accept positive and negative introspection for (implicit) belief:

(**4**) $\quad \mathsf{B}\,A \rightarrow \mathsf{B}\,\mathsf{B}\,A$

(**5**) $\quad \neg\,\mathsf{B}\,A \rightarrow \mathsf{B}\,\neg\,\mathsf{B}\,A$

The logic that results by adding the schemas **D**, **4**, and **5** to the axiomatic basis for K is known as KD45.

Exercise 5.12

Is a transitive, serial, and euclidean relation always symmetric? If yes, explain why. If no, give a counterexample. What does your result mean for the validity of principle **B** in KD45?

Exercise 5.13

Show (in any way you like) that $\mathsf{B}(\mathsf{B}\,A \to A)$ is valid if the logic of belief is KD45.

If we want to model the connection between knowledge and belief, we need a multi-modal language that has both a K operator and a B operator. Models for this language will have two accessibility relations (for each agent), one for knowledge, the other for belief.

The power of combined logics for knowledge and belief lies in the interaction principles that plausibly link the two concepts. Here is a list of popular principles that don't follow from the individual logics of knowledge and belief.

(**KB**) $\mathsf{K}\,A \to \mathsf{B}\,A$

(**PI**) $\mathsf{B}\,A \to \mathsf{K}\,\mathsf{B}\,A$

(**NI**) $\neg\,\mathsf{B}\,A \to \mathsf{K}\,\neg\,\mathsf{B}\,A$

(**SB**) $\mathsf{B}\,A \to \mathsf{B}\,\mathsf{K}\,A$

KB assumes that knowledge implies belief. **PI** and **NI** strengthen the introspection principles for belief, assuming that agents always (implicitly) know what they believe or disbelieve. **SB** assumes that if an agent believes something, then they also believe that they know it. This is sometimes said to reflect a conception of "strong belief", on which belief is incompatible with doubt. If you believe p in the sense that you have no doubt that p, then you plausibly believe that you know p.

In the previous section, I used some of these principles to argue that K does not validate **5**. The argument went something like this.

1. Assume the **T**-schema is not valid for belief. So there are conceivable scenarios in which $\mathsf{B}\,p$ is true and p false (on some interpretation of p).
2. By **SB** (and *modus ponens*), it follows that $\mathsf{B}\,\mathsf{K}\,p$ is true (in these scenarios).
3. By the **D**-schema for belief, we can infer $\neg\,\mathsf{B}\,\neg\,\mathsf{K}\,p$.
4. By **KB**, $\mathsf{K}\,\neg\,\mathsf{K}\,p$ entails $\mathsf{B}\,\neg\,\mathsf{K}\,p$. Since the latter is false, we have $\neg\,\mathsf{K}\,\neg\,\mathsf{K}\,p$.
5. From $\neg p$, we also have $\neg\,\mathsf{K}\,p$, by the **T**-schema for knowledge.
6. So if the **T**-schema is not valid for belief, then $\neg\,\mathsf{K}\,p$ does not entail $\mathsf{K}\,\neg\,\mathsf{K}\,p$.

More interestingly, the above interaction principles, together with the **D**-schema for belief, imply that an agent believes a proposition just in case she doesn't know that she doesn't know it:

(**BMK**) $\mathsf{B}\,A \leftrightarrow \mathsf{M}\,\mathsf{K}\,A$

So belief is definable in terms of knowledge.

Here is how we can get from $\mathsf{B}\,A$ to $\mathsf{M}\,\mathsf{K}\,A$.

1. Suppose $\mathsf{B}\,A$.
2. By **SB**, it follows that $\mathsf{B}\,\mathsf{K}\,A$.
3. By **D**, it follows that $\neg\,\mathsf{B}\,\neg\,\mathsf{K}\,A$.
4. By **KB**, it follows that $\neg\,\mathsf{K}\,\neg\,\mathsf{K}\,A$, and so that $\mathsf{M}\,\mathsf{K}\,A$.

To show that $\mathsf{M}\,\mathsf{K}\,A$ entails $\mathsf{B}\,A$, I'll show that $\neg\,\mathsf{B}\,A$ entails $\neg\,\mathsf{M}\,\mathsf{K}\,A$.

1. By **KB**, $\neg\,\mathsf{B}\,A \rightarrow \neg\,\mathsf{K}\,A$ is a logical truth.
2. Since logical truths are true at every world, we have $\mathsf{K}(\neg\,\mathsf{B}\,A \rightarrow \neg\,\mathsf{K}\,A)$.
3. By the **K**-schema, it follows that $\mathsf{K}\,\neg\,\mathsf{B}\,A \rightarrow \mathsf{K}\,\neg\,\mathsf{K}\,A$.
4. Now suppose $\neg\,\mathsf{B}\,A$.
5. By **NI**, it follows that $\mathsf{K}\,\neg\,\mathsf{B}\,A$.
6. By 3 above, it follows that $\mathsf{K}\,\neg\,\mathsf{K}\,A$, which is equivalent to $\neg\,\mathsf{M}\,\mathsf{K}\,A$.

Given the equivalence between $\mathsf{B}\,A$ and $\mathsf{M}\,\mathsf{K}\,A$, the **D**-schema for belief

$$\mathsf{B}\,A \rightarrow \neg\,\mathsf{B}\,\neg A$$

is equivalent to

$$\mathsf{M}\,\mathsf{K}\,A \rightarrow \neg\,\mathsf{M}\,\mathsf{K}\,\neg A$$

which in turn is equivalent to

$$\mathsf{M}\,\mathsf{K}\,A \rightarrow \mathsf{K}\,\mathsf{M}\,A.$$

This is the **G** schema for knowledge. So if we accept the above interaction principles, and principle **D** for belief, then the logic of knowledge must validate **G**.

Exercise 5.14

Show that the interaction principles entail principles **4** and **5** for belief.

Exercise 5.15

Suppose the logic of knowledge validates **5**, the logic of belief validates **D**, and we have the interaction principles **KB** and **SB**. Show that knowledge is then equivalent to belief: $KA \leftrightarrow BA$ comes out as valid. (Another reason to think that that **5** is not valid in the logic of knowledge.)

Exercise 5.16

There seems to be no natural expression in English for the dual of belief. A common way to express that someone does not believe not p is to say that they believe that it might be that p, which seems to have the surface form $\Box\Diamond p$. Explain why, if the logic of belief is KD45, then $\Box\Diamond p$ is equivalent to the dual of $\Box$.

It can also be instructive to combine epistemic with non-epistemic operators. Philosophers have often been interested not just in what we *do* know, but also in what we *can* know. Various skeptical arguments suggest that we *cannot know* that we have hands. For another example, the "verificationist" movement of the 20th century assumed that a sentence is meaningful only if its truth-value can in principle be settled by mathematical proof or empirical investigation; in other words, a sentence is meaningful only if *it is possible to know* that it is true.

We can formalize claims like these in a multi-modal language with a diamond $\Diamond$ for 'in principle possible' and a knowledge operator K for 'someone knows'. The verificationist hypothesis that every truth is in principle knowable can then be expressed by the following interaction principle:

(Knowability) $A \rightarrow \Diamond \mathsf{K} A$

The following neat little argument, due to Alonzo Church, shows that this interaction principle is untenable.

1. Let p be any unknown truth. (Nobody thinks all truths are actually known.)
2. So we have $p \wedge \neg \mathsf{K}\, p$.
3. By the Knowability principle, it follows that $\Diamond \mathsf{K}(p \wedge \neg \mathsf{K}\, p)$.
4. By the **K**-schema for knowledge, $\mathsf{K}(p \wedge \neg \mathsf{K}\, p)$ entails $\mathsf{K}\, p \wedge \mathsf{K} \neg \mathsf{K}\, p$.
5. By the **T**-schema for knowledge, $\mathsf{K} \neg \mathsf{K}\, p$ entails $\neg \mathsf{K}\, p$.
6. So $\mathsf{K}(p \wedge \neg \mathsf{K}\, p)$ entails both $\mathsf{K}\, p$ and $\neg \mathsf{K}\, p$.
7. So $\mathsf{K}(p \wedge \neg \mathsf{K}\, p)$ is logically impossible.
8. So $\neg \Diamond \mathsf{K}(p \wedge \neg \mathsf{K}\, p)$.
9. This contradicts the application of the Knowability principle on line 3.

Exercise 5.17

Show that if the logic of belief is at least KD4, then there are *unbelievable truths*: truths of which it is impossible that anyone believes them. (You can assume that there are truths which no-one in fact believes.)