5 Epistemic Logic

5.1 Epistemic accessibility

When we say that something is possible, we often mean that it is compatible with our information. This "epistemic" flavour of possibility – along with related concepts such as knowledge, belief, information, and communication – is studied in epistemic logic.

Standard epistemic logic relies heavily on the possible-worlds semantics introduced in chapters 2 and 3. The guiding idea is that *information rules out possibilities*. Imagine we are investigating a crime. There are three suspects: the gardener, the butler, and the cook. Now a credible eye-witness tells us that the gardener was out of town at the time of the crime. This allows us to rule out the previously open possibility that the gardener is the culprit. When we gain information, the space of open possibilities shrinks.

Let's say that a world is *epistemically accessible* for an agent if it is compatible with the agent's knowledge. Recall that a world is a maximally specific possibility. For any such possibility, we may ask whether it might be the actual world. If our information allows us to give a negative answer then the world is not epistemically possible for us – it is epistemically inaccessible. Before we learned that the gardener was out of town, our epistemically accessible worlds included worlds at which the gardener committed the crime. When we received the eye-witness report, these worlds became inaccessible.

Exercise 5.1

Which worlds are epistemically accessible for an agent who knows all truths?
Which worlds are epistemically accessible for an agent who knows nothing?

We will interpret the box and the diamond in terms of epistemic accessibility. In

this context, the box is usually written ‘K’. For once, this doesn’t stand for Kripke but for knowledge. I will use ‘M’ (‘might’) for the diamond. So $\mathsf{K}A$ means that A is true at all epistemically accessible worlds, while $\mathsf{M}A$ means that A is true at some epistemically accessible world. If we want to clarify which agent we have in mind, we can add a subscript: $\mathsf{M}_\mathsf{b}A$ might say that A is epistemically possible for Bob.

We often informally read K as ‘the agent knows’. In at least one respect, however, our K operator does not match the knowledge operator of ordinary English.

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. For example, if $p \to q$ and p are both true at w, then so is q (by definition 3.2). As a consequence, if $p \to q$ and p are true at all epistemically accessible worlds (for some agent), then q is also true at all these worlds. $\mathsf{K}(p \to q)$ and $\mathsf{K}p$ together entail $\mathsf{K}q$. More generally, the K operator is **closed under logical consequence**, meaning that if B logically follows from $A_1, \dots, A_n$, and $\mathsf{K}A_1, \dots, \mathsf{K}A_n$, then $\mathsf{K}B$.

Our ordinary conception of knowledge does not seem to be closed under logical consequence. If you know the axioms of a mathematical theory, you don’t automatically know everything that logically follows from the axioms. Our K operator might be taken to formalise the concept of *implicit knowledge*, where an agent implicitly knows a proposition if the proposition follows from things the agent knows. An agent’s implicit knowledge represents the information the agent has about the world. If what you know entails p, then the information you have settles that p, even though you may not realise that it does.

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) Alice knows whether it is raining.
(d) You know that you’re guilty if you don’t know that you’re innocent.

5.2 The logic of knowledge

What is the logic of (implicit) knowledge? Which sentences in the language of epistemic logic are valid? Which are logical consequences of which others?

The basic system K is arguably too weak. There are Kripke models in which $\Box p$ is true at some world while p is false. But knowledge entails truth. If p is genuinely known (or entailed by what is known) then p is true. In the logic of knowledge, all instance of the (T)-schema are valid.

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

We know from section 3.4 that the (T)-schema corresponds to reflexivity, in the sense that all instances of the schema are valid on a frame iff the frame is reflexive. To ensure that all (T) instances are valid, we will therefore assume that Kripke models for epistemic logic are always reflexive. Every world is accessible from itself.

This makes sense if you remember what accessibility means in epistemic logic. We said that a world v is (epistemically) accessible from a world w if v is compatible with what the agent knows at w. Whatever the agent knows at w must be true at w. So any world in any conceivable scenario must be accessible from itself.

Let's look at other properties of the epistemic accessibility relation. Is the relation symmetric? If v is compatible with what is known at w, is w compatible with what is known at v? I will give two arguments for a negative answer.

My first argument assumes that we have non-trivial knowledge about the external world. Let's say we know that we have hands. Now consider a possible world in which we are brains in a vat, falsely believing that we have hands. In that world, we know very little. We don't know that we have hands, nor that we are handless brains in a vat. Perhaps we know that we are conscious, and what kinds of experiences we have. But since our experiences are the same in the vat world and in the actual world (let's assume), the actual world is compatible with what little we know in the vat world. So the actual world is accessible from the vat world. But the vat world is not accessible from the actual world – otherwise we wouldn't know that we have hands. If the actual world is accessible from the vat world and the vat world is inaccessible from the actual world then the accessibility relation isn't symmetric.

My second argument starts with a scenario in which someone has misleading evidence that some proposition p is false. This is easily conceivable. In that scenario,

p is true but the agent believes $\neg p$. Often, when we believe something, we also believe that we know it. Let's assume that our agent believes that they know $\neg p$. Let's also assume that their beliefs are consistent, so they don't believe that they *don't* know $\neg p$. Since they don't believe this proposition (that they don't know $\neg p$) they don't know it either: they don't know that they don't know $\neg p$. So we have a scenario in which p is true but $\mathsf{K}\neg\mathsf{K}\neg p$ false.

Can you see what this has to do with symmetry? In section 3.4 I mentioned that symmetry corresponds to the schema

(B) $\quad A \to \mathsf{K}\,\mathsf{M}\,A.$

This means that all instances of (B) are valid on a frame iff the frame is symmetric. If the epistemic accessibility relation were symmetric, then all instances of (B) would be valid. But I've just described a scenario in which an instance of (B) is false. So the epistemic accessibility relation isn't symmetric.

What about transitivity, which corresponds to schema (4)?

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

In epistemic logic, (4) is known as the **KK principle**, or (misleadingly) as **positive introspection**. There is an ongoing debate over whether the principle should be considered valid. I will review one argument for either side.

A well-known argument against the KK principle draws on the idea that knowledge requires "safety": you know p only if you couldn't easily have been wrong about p. To motivate this idea, consider a Gettier case. 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 genuinely 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 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's belief state *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 this means), and let $\Box$ be a non-epistemic operator that formalizes 'at all worlds'. $\Box(N \to A)$ then means that A is true at all world at which conditions are normal. According to the definition I just gave, a belief state s indicates p iff

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

The state s indicates that s indicates p iff

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

A quick tree proof reveals that (*) entails (**). That is, whenever a state indicates p then 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 5.3

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

The (4)-schema says that people have knowledge of their knowledge. The (5)-schema says that people have knowledge of their ignorance: if you don't know something, then you know that you don't know it. This hypothesis is (misleadingly)

known as **negative introspection**.

(5) $\quad \mathsf{M}A \to \mathsf{K}\mathsf{M}A.$

We know that the (5)-schema corresponds to euclidity. This gives us a quick argument against the schema. As you showed in exercise 3.3, reflexivity and euclidity together entail symmetry. The epistemic accessibility relation is reflexive. If it were euclidean, it would be symmetric. But I've argued that it isn't symmetric. So the logic of knowledge doesn't validate (5).

We can also give a more direct argument against negative introspection. Consider again a scenario in which someone has misleading evidence that some proposition p is false. Since p is actually true, the agent doesn't know $\neg p$. But the agent might not know that they don't know $\neg p$. (On the contrary, they might believe that they do know $\neg p$.) In that scenario, $\neg\mathsf{K}\neg p$ is true but $\mathsf{K}\neg\mathsf{K}\neg p$ is false.

Here it is important to not be misled by a curiosity of ordinary language. When we say that someone doesn't know p, this seems to imply that p is true. 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. You might therefore be tempted to regard all instances of the following schema as valid:

(NT) $\quad \neg\mathsf{K}A \to A$

On reflection, however, (NT) is unacceptable. If $\neg\mathsf{K}A$ entails A, then by contraposition $\neg A$ entails $\mathsf{K}A$: everything that is false would be known! Indeed, if I *don't* have a pet aardvark then surely my neighbour does not know that I have one. We shall therefore not regard the inference from $\neg\,\mathsf{K}A$ to A as valid.

Exercise 5.4

Can you find a Kripke frame on which (NT) is valid?

Exercise 5.5

Let's say that an agent is *ignorant of* a proposition if they don't know the proposition and the proposition is true. (In English, saying that someone doesn't know a proposition normally conveys that they are ignorant of the proposi-

tion, in this sense.) Show that if the logic of knowledge is at least as strong as K, then ignorance of A entails ignorance of ignorance of A.

We have looked at six schemas: (T), (B), (4), (5), and (NT). Philosophers working in epistemic logic generally reject (B), (5), and (NT), accept (T), and are divided over (4). Theorists in other disciplines often assume that the logic of knowledge is S5, which would render all instances of (T), (4), (B), and (5) valid. If we drop (B) and (5) but keep (T) and (4), we get S4. If we also drop (4), we get system T.

But we might look at other schemas, corresponding to further conditions on the accessibility relation. For example, some have argued that we should adopt a weakened form of negative introspection. The above counterexample to negative introspection – schema (5) – involved an agent who doesn't know that they don't know a certain proposition because they don't know that the proposition is false. This kind of counterexample can't arise if the relevant proposition is true. One might therefore suggest that if an agent doesn't know a proposition *p and p is true*, then the agent always knows that they don't know *p*. This would give us a schema known as 0.4:

$$(0.4) \qquad (\neg \mathsf{K} A \wedge A) \rightarrow \mathsf{K} \neg \mathsf{K} A$$

All instances of (0.4) are S5-valid, but not all of them are S4-valid. Adding the schema to S4 leads to a system known as S4.4.

Exercise 5.6

Explain why Gettier cases cast doubt on (0.4).

A more modest extension of S4 adds the schema (G), which corresponds to convergence of the accessibility relation:

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

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 section 5.4.

Exercise 5.7

Use the tree method to check the following claims. (See the table at the end of chapter 3 for the tree rules that go with B, S4, and S4.2.)

(a) $\models_T \mathsf{M}\mathsf{K}p \to \mathsf{K}\mathsf{M}p$.
(b) $\models_B \mathsf{M}\mathsf{K}p \to \mathsf{K}\mathsf{M}p$.
(c) $\models_{S4} \mathsf{M}\mathsf{K}\mathsf{M}p \to \mathsf{M}p$.
(d) $\models_{S4} \mathsf{M}\mathsf{K}p \leftrightarrow \mathsf{K}\mathsf{K}p$.
(e) $\models_{S4} \mathsf{M}\mathsf{K}(p \to \mathsf{K}\mathsf{M}p)$.
(f) $\models_{S4.2} (\mathsf{M}\mathsf{K}p \wedge \mathsf{M}\mathsf{K}q) \to \mathsf{M}\mathsf{K}(p \wedge q)$.

5.3 Multiple Agents

A world that is epistemically accessible for one agent may not be accessible for another. If we want to reason about the information available to different agents, we need separate K operators and accessibility relations for each agent.

We can easily expand the language $\mathfrak{L}_M$ to a **multi-modal language** by introducing a whole series of box operators $\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_3, \ldots$ with their duals $\mathsf{M}_1, \mathsf{M}_2, \mathsf{M}_3, \ldots$. This multi-modal language is interpreted in multi-modal Kripke models.

Definition 5.1

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

- a non-empty set W,
- a set of binary relation $R_1, R_2, R_3, \ldots$ on W, and
- a function V that assigns to each sentence letter a subset of W.

In our present application, every accessibility relation R_i represents what information is available to a particular agent. A world v is R_i-accessible from w iff v is compatible with the information agent i has at world w.

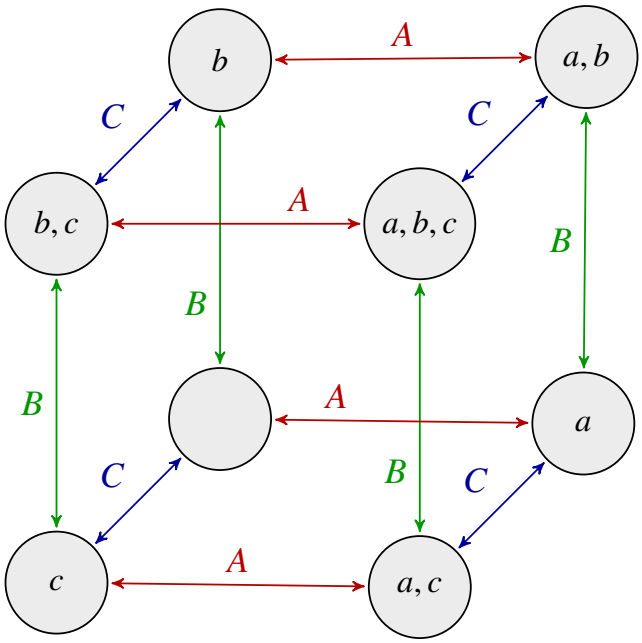
The definition of truth at a world in a Kripke model (definition 3.2) is easily extended to multi-modal Kripke models. Instead of clauses (g) and (h), we have the following conditions, for each pair of a modal operator $\mathsf{K}_i/\mathsf{M}_i$ and the corresponding accessibility relation R_i:

$M, w \models \mathsf{K}_i A$ iff $M, v \models A$ for all v in W such that wR_iv.
$M, w \models \mathsf{M}_i A$ iff $M, v \models A$ for some v in W such that wR_iv.

As an application of this machinery, let's look at 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 have mud on their face. As they come 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 that they don't know. Mother asks again, 'Do you know if you have mud on your face?'. This time, two children say that they know. Do you know many children have mud on their face? What happens when the mother asks her question a third time?

To answer these questions, we can begin by drawing a model. I'll call the three children Alice, Bob, and Carol, and I'll use a, b, c as sentence letters expressing, respectively, that Alice/Bob/Carol is muddy. Before the mother's first announcement, there are eight relevant possibilities.

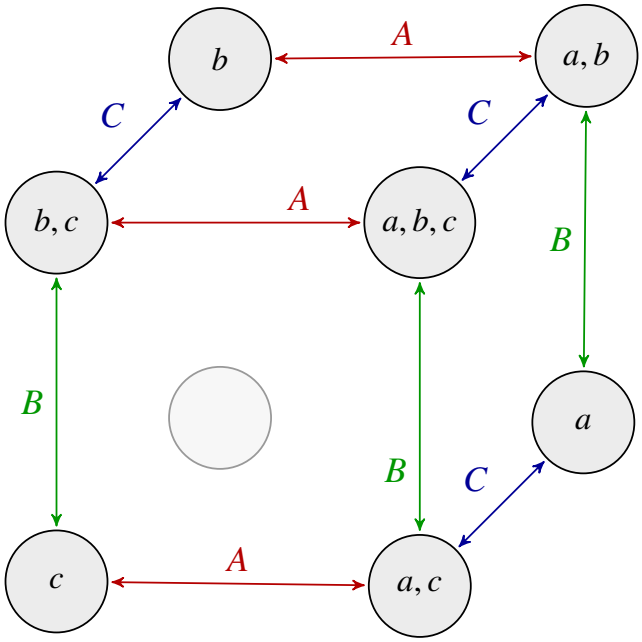


Since we have three epistemic agents, we have three accessibility relations, one for Alice (drawn in red), one for Bob (green), and one for Carol (blue). To remove clutter, I have left out the (3 × 8) arrows leading from each world to itself, but we should keep in mind that every world is also accessible from itself, for each agent.

Don't confuse an arrow in the diagram of a model with an accessibility relation. We have three accessibility relations, but more than three arrows. All the red arrows in the picture represent one and the same accessibility relation. The accessibility relation for Alice holds between a world and another whenever a red arrow leads from the first world to the second.

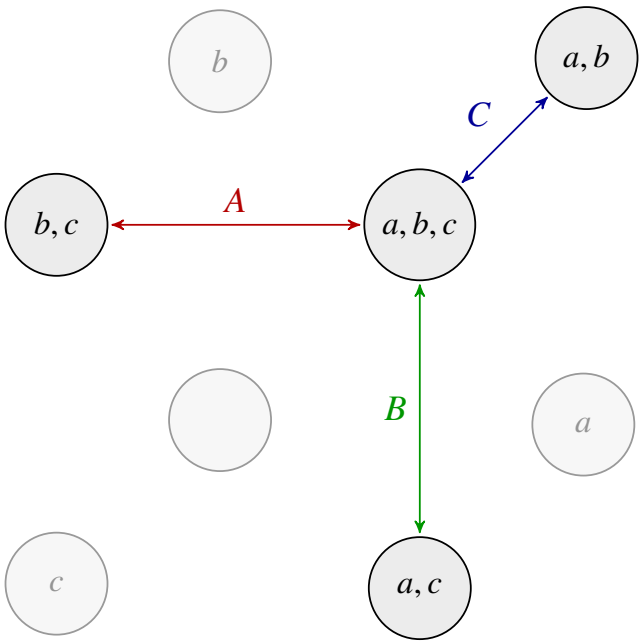
Notice how the fact that every child can see the others is reflected in the diagram. For example, at the top left world, where only Bob is muddy (b), Alice sees that Bob is muddy and that Carol is clean; the only epistemic possibilities for Alice at that world are the two worlds at the top: the b world itself and the a, b world to the right. 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 world where a, b, and c are all false. More importantly, it allows *each child* to rule out the this 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 (b) Alice doesn't know whether she is muddy, because the a, b world (top right) is A-accessible; nor does Carol know whether she is muddy, because the b, c world is C-accessible. But Bob knows that he is muddy: no other world is B-accessible. Intuitively, at the b 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 (a, b, c), however, no child knows whether they are muddy. That world is not actual, and it is no longer accessible for

anyone. The remaining open possibilities are the b, c world, the a, c world, and the a, b world, each of which is only accessible from itself.

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

Exercise 5.8

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:

5 May, 6 May, 9 May
7 June, 8 June
4 July, 6 July
4 August, 5 August, 7 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?

What logic do we have for our multi-modal language? Each pair of a K_i and M_i operator should obey whatever conditions we want to impose on the logic of knowledge. Are there also new principles governing the interaction between operators for different agents?

We plausibly want all instances of the following to come out valid:

$$\mathsf{K}_1 \mathsf{K}_2 A \to \mathsf{K}_1 A.$$

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

A common assumption in epistemic logic is that there are no genuinely new in-

teraction principles for the knowledge of multiple agents – no principles that don't already follow from the logic of individual knowledge. The above principle, for example, is entailed by the assumption that the (T)-schema holds for K_2. 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 holds for K_2, then every world is R_2-accessible from itself. In particular, then, any R_1-accessible world is R_2-accessible from itself. It follows that A holds at every R_1-accessible world. So $\mathsf{K}_1\,A$ is true.

We can use the tree rules to streamline arguments like this. When multiple agents are in play, we need to keep track of which world is accessible for which agent. When expanding a node of type $\mathsf{M}_i\,A\ \ (w)$, for example, we add a node wR_iv, with subscript i, and another node $A\ \ (v)$.

Here is a tree proof of the schema $\mathsf{K}_1\,\mathsf{K}_2\,A \to \mathsf{K}_1\,A$, assuming that R_2 is reflexive.

1.	$\neg(\mathsf{K}_1\,\mathsf{K}_2\,A \to \mathsf{K}_1\,A)$	(w)	(Ass.)
2.	$\mathsf{K}_1\,\mathsf{K}_2\,A$	(w)	(1)
3.	$\neg\,\mathsf{K}_1\,A$	(w)	(1)
4.	wR_1v		(3)
5.	$\neg A$	(v)	(3)
6.	$\mathsf{K}_2\,A$	(v)	(2,4)
7.	vR_2v		(Refl.)
8.	A	(v)	(6,7)
	×		

Exercise 5.9

Use the tree method to check which of the following interaction principles are valid if the logic of individual knowledge is S4. If a principle is invalid, give a counterexample.

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

We can also define new modal operators for groups of agents. A proposition is said to be **mutually known** in a group G if 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 (at least for finite groups). 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), it does not follow that E_G validates (4). For a counterexample, consider a group of two agents; both know p, and both know of themselves that they know p, 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 5.10

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 has proved useful in many areas 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 $\mathsf{K}_1, \ldots, \mathsf{K}_n$. Still, we can define it semantically in terms of the accessibility relations for the individual agents: $\mathsf{C}_G A$ 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 the agents' accessibility relations.

It is easy to see that common knowledge validates (all instances of) (4). It validates (T) whenever individual knowledge validates (T). So the logic of common knowledge is at least S4. The complete logic of common knowledge also contain some 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) $\quad (A \wedge \mathsf{C}_G(A \rightarrow \mathsf{E}_G A)) \rightarrow \mathsf{C}_G A$

You may want to confirm that these are valid. (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.

An agent's belief state represents the world as being a certain way. For every possible world, we can ask whether it matches what the agent believes. If, for example, your only non-trivial belief is that there are seventeen types of parrot, then every world in which there are seventeen types of parrot matches your beliefs. Every such world is *doxastically accessible* for you. As you acquire further beliefs, the space of doxastically accessible worlds becomes smaller and smaller.

We interpret $\mathsf{B}\,p$ as saying that p is true at all doxastically accessible worlds (for the agent we have in mind). Since we won't spend a lot of time with this operator, we will simply write its dual as $\neg\,\mathsf{B}\,\neg$.

The logic of B is different from the logic of K, if only because beliefs can be false. So we will not regard all instances of

(T) $\quad \mathsf{B}\,A \to A$

as valid. We may, however, accept the weaker schema

(D) $\quad \mathsf{B}\,A \to \neg\,\mathsf{B}\,\neg A.$

This reflects the assumption that a belief state that represents the world as being a certain way A can't also represent the world as being the opposite way $\neg A$.

In the previous section, I argued that (implicit) knowledge does not validate the negative introspection principle (5), and I reviewed an argument against the positive introspection principle (4). Neither argument carries over to belief. Many epistemic logicians accept positive and negative introspection for (implicit) belief:

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

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

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

Exercise 5.11

Is a transitive, serial, and euclidean relation always symmetric? If yes, explain why. If no, give a counterexample. What does your result mean for schema (B) in KD45?

Exercise 5.12

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 with both the K operator and the B operator. Models for this language will have two accessibility relations R_e and R_d. The first represents epistemic accessibility and is used for the interpretation of K, the second represents doxastic accessibility and is used to interpret B.

The power of combined logics for (implicit) knowledge and belief lies in the interaction principles that might 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. They assume that a state of belief or disbelief is always known to the agent. (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.

These interaction principles, together with the (D)-schema for belief, imply that

an agent believes a proposition just in case they don't know that they don't know it:

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

Somewhat surprisingly, then, we could define belief 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).

(In fact, we don't need to assume that the interaction principles and (D) hold for our ordinary concept of belief. As long as one can coherently define a concept B that validates these principles we can derive the (G)-schema for K.)

Exercise 5.13

Show that the interaction principles entail principles (4) and (5) for belief: $\mathsf{B}\,A \to \mathsf{B}\,\mathsf{B}\,A$ and $\neg\mathsf{B}\neg A \to \mathsf{B}\neg\mathsf{B}\neg A$.

Exercise 5.14

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 (5) is not valid in the logic of knowledge.)

Exercise 5.15

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 has the surface form $\Box\Diamond p$. Can you explain why this might be an adequate way of expressing $\Diamond p$?

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, for example, suggest that we *cannot know* that we have hands. For another example, the "verificationist" movement in the early 20th century assumed that a sentence is meaningful only if its truth-value can in principle be settled by mathematical proof or empirical investigation. This would imply that 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 knowledge operator K and a diamond $\Diamond$ for the relevant kind of circumstantial possibility. The verificationist hypothesis that every truth is in principle knowable is then expressed by the following interaction principle:

(Knowability) $A \to \Diamond\,\mathsf{K}\,A$

The principle is refuted by the following argument, due to Alonzo Church.

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. In any logic that extends the minimal system K, $\mathsf{K}(p \wedge \neg \mathsf{K} p)$ entails $\mathsf{K} p \wedge \mathsf{K} \neg \mathsf{K} p$.
4. By the (T)-schema for knowledge, $\mathsf{K} \neg \mathsf{K} p$ entails $\neg \mathsf{K} p$.
5. So $\mathsf{K}(p \wedge \neg \mathsf{K} p)$ entails both $\mathsf{K} p$ and $\neg \mathsf{K} p$.
6. So the hypothesis $\mathsf{K}(p \wedge \neg \mathsf{K} p)$ is inconsistent.
7. So $\neg \Diamond \mathsf{K}(p \wedge \neg \mathsf{K} p)$.
8. Lines 2 and 7 together provide a counterexample to the Knowability principle.

Exercise 5.16

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.)