Logic 2: Modal Logic

Wolfgang Schwarz

September 1, 2019

© 2019 Wolfgang Schwarz
www.wolfgangschwarz.net/logic2

This work is licensed under a Creative Commons "Attribution-NonCommercial-ShareAlike 4.0 International" license.

Contents

1 Modal Operators

1.1 Boxes and diamonds

Modal logic is an extension of propositional and predicate logic that is widely used (across many disciplines) to reason about possibility and necessity, obligation and permission, the flow of time, the processing of computer programs, and a range of other topics. Each of these applications begins by adding new symbols to the formal language of classical propositional or predicate logic. Before we explore such additions, let's briefly review why we use formal languages in the first place.

When reasoning about a given topic, we want to make sure that the stated conclusions really follow from the stated premises. If they do, we say that the reasoning is valid. A little more precisely, an argument is **valid** if there is no conceivable situation in which the premises are true while the conclusion is false.

Here is an example of a valid argument.

> All myriapods are oviparous.
> Some arthropods are myriapods.
> Therefore: some arthropods are oviparous.

You can tell that this argument is valid even if you don't understand the zoological terms, because every argument of the same **logical form** is valid. The logical form of the above argument might be expressed as follows.

> All F are G.
> Some H are F.
> Therefore: some H are G.

No matter what descriptive terms you plug in for F, G, and H, you get a valid argument. The argument about myriapods is therefore not just valid, but **logically valid** – valid in virtue of its form.

In natural languages like English, the logical form of sentences is not always transparent. 'All dogs barked at a tree' can mean either that there is a single tree at which all dogs barked, or that each dog barked at some tree or other. The two readings have different logical consequences, so it would be good to keep them apart. Also, the meaning of logical expressions ('all', 'some', 'and', etc.) in natural language is often unclear and complicated. 'Paul and Paula got married and had children' suggests that the marriage came before the children. In 'Paul went to the zoo and Paula stayed at home', the word 'and' does not seem to have this temporal meaning.

To get around these problems, we invent formal languages whose logical expressions have precise meanings and in which there are no ambiguities of logical form. If we want to evaluate natural-language arguments for logical validity, we first have to translate them into the formal language. (Sometimes an argument will be valid on one translation and invalid on another.) We can also reason directly in the formal language – perhaps to show that under a certain interpretation of the logical terms, a certain hypothesis logically follows from certain assumptions.

Returning to modal logic, consider the following argument.

> It might be raining.
> There is no doubt that we will get wet if it is raining.
> Therefore: we might get wet.

The argument looks valid. Indeed, any argument of this form is plausibly valid:

> It might be that A.
> There is no doubt that B if A.
> Therefore: it might be that B.

But it's hard to bring out the validity of these arguments in classical propositional or predicate logic. We need formal expressions corresponding to 'it might be that' and 'there is no doubt that'. The language of classical propositional or predicate logic does not have such expressions.

So let's add them. That is, let's invent a new formal language with two new logical symbols. It doesn't matter what these look like; a popular choice in modal logic is to use a diamond $\Diamond$ and a box $\Box$. (We will use various other symbols in later chapters.) Let's assume that the diamond formalizes 'it might be that', and the box 'there is no

doubt that' or (equivalently) 'it is certain that'. We can then formalize the above argument as follows.

$$\frac{\Diamond r \qquad \Box(r \rightarrow w)}{\Diamond w}$$

Of course, merely adding new symbols doesn't help much. We also need to lay down some new rules for how to reason with these symbols. The rules should be motivated by what the symbols are supposed to mean. So we shall also assign a more precise meaning to the diamond and the box – just as classical logic assigns a precise meaning to the symbol $\wedge$ which may or may not exactly match the meaning of 'and' in English. We can then confirm that no matter what r and w mean, if $\Diamond r$ and $\Box(r \rightarrow w)$ are both true, then so is $\Diamond w$.

We have entered the realm of modal logic.

Historically, modal logic grew out of the study of necessity and contingency, which medieval logicians regarded as "modes of truth"; hence the name 'modal logic'. Today, the study of necessity and contingency is but one of many subfields within modal logic. To a first approximation, any part of logic that involves *non-truth-functional sentence operators* is part of modal logic.

By a **sentence operator** I mean an expression that combines with one or more sentences to create a new sentence. Classical propositional logic has the sentence operators $\neg$ (not), $\wedge$ (and), $\vee$ (or), $\rightarrow$ (if-then), and $\leftrightarrow$ (if-and-only-if). In modal logic, we have further operators such as $\Diamond$ and $\Box$. (Sentence operators are also called 'connectives'.)

The sentence operators of classical propositional logic are all truth-functional. Remember that an operator is **truth-functional** if the truth-value of a complex sentence built with the use the operator is determined by the truth-value of its parts. For example, a conjunction $A \wedge B$ is true whenever A and B are both true, and false otherwise. If you know the truth-value of the conjuncts, you know the truth-value of the conjunction.

Things are different for our operators $\Diamond$ and $\Box$. The truth-value of $\Diamond A$ and $\Box A$ is *not* determined by the truth-value of A. That is, there are conceivable scenarios in which two sentences A and B have the same truth-value whereas $\Diamond A$ and $\Diamond B$ (or $\Box A$ and $\Box B$) have different truth-values.

For example, let r translate 'it is currently raining in Sydney', and let t translate '2+2=5'. Reading the diamond as 'it might be the case that', it is easy to imagine a scenario in which $\Diamond r$ and $\Diamond \neg r$ are both true, while $\Diamond t$ is false. Since r and $\neg r$ have opposite truth values, one of them must be false. If r is false, then r and t have the same truth-value while $\Diamond r$ and $\Diamond t$ have different truth-values; if $\neg r$ is false, then $\neg r$ and t have the same truth-value while $\Diamond \neg r$ and $\Diamond t$ have different truth-values. Either way, we have a counterexample to the truth-functionality of the diamond.

Exercise 1.1

Which of these English expressions are truth-functional?

(a) It used to be the case that . . .
(b) It is widely known that . . .
(c) It is false that . . .
(d) It is necessary that . . .
(e) I can see that . . .
(f) God believes that . . .
(g) Either 2+2=4 or it is practically feasible that . . .

The meaning of truth-functional operators can be specified by a truth-table. All you need to know to understand $\wedge$ is that $A \wedge B$ is true whenever A and B are both true (and false otherwise). The meaning of $\Diamond$ and $\Box$, by contrast, cannot be given by a truth-table. The standard approach to define the meaning of modal operators instead involves the concept of possible worlds. Roughly, we'll interpret $\Diamond A$ as saying that A is true at some possible world, and $\Box A$ as saying that A is true at all possible worlds.

Much more on this later.

1.2 Reasoning with boxes and diamonds

Let's be clear about our formal language. If we add the box and the diamond to the language of classical propositional logic, we get the **standard language of modal propositional logic**, for short, $\mathfrak{L}_M$. We will stick with propositional logics until chapter 9, when we turn to modal predicate logic.

The sentences of $\mathfrak{L}_M$ are defined as follows.

1. Every sentence letter $p, q, r, \ldots$ is an $\mathfrak{L}_M$-sentence.
2. If A is an $\mathfrak{L}_M$-sentence, then so are $\neg A$, $\Diamond A$, and $\Box A$.
3. If A and B are $\mathfrak{L}_M$-sentences, then so are $(A \wedge B)$, $(A \vee B)$, $(A \rightarrow B)$ and $(A \leftrightarrow B)$.
4. Nothing else is an $\mathfrak{L}_M$-sentence.

As usual, outermost parentheses are omitted when displaying sentences. So $p \wedge q$ is treated as an abbreviation of $(p \wedge q)$.

Exercise 1.2

Which of these are $\mathfrak{L}_M$-sentences?

(a) p
(b) $\Diamond$
(c) $\Diamond p \vee (\Box p \rightarrow p)$
(d) $\Box\Box p$
(e) $\Box A \rightarrow A$
(f) $(\Diamond r \wedge \Diamond q r) \wedge \Diamond\Box\Diamond\Box p$

The sentence letters of $\mathfrak{L}_M$ don't have a determinate meaning. It doesn't make sense to ask, without further information, whether p or $\Diamond q$ are true sentences of $\mathfrak{L}_M$. However, it does make sense to ask whether a given $\mathfrak{L}_M$-sentence logically follows from others, for this only depends on the logical form of the relevant sentences.

For example, in the previous section I claimed that if the diamond formalizes 'it might be that' and the box 'there is no doubt that', and if r and w translate 'it is raining' and 'we will get wet' respectively, then $\Diamond r$ and $\Box(r \rightarrow w)$ logically entail $\Diamond w$. But if the entailment is genuinely a matter of logic, then it doesn't depend on the meaning of r and w. No matter what meaning we give to r and w, $\Diamond r$ and $\Box(r \rightarrow w)$ will entail $\Diamond w$.

Logicians often use the symbol '$\models$' (the "double-barred turnstile") to express logical consequence. The claim that $\Diamond r$ and $\Box(r \rightarrow w)$ logically entail $\Diamond w$ can then be expressed as follows:

$$\Diamond r, \Box(r \rightarrow w) \models \Diamond w.$$

Intuitively, this says that there is no conceivable scenario in which $\Diamond r$ and $\Box(r \rightarrow w)$ are true while $\Diamond w$ is false, no matter what meaning is given to r and w.

Note that '$\models$' is not a symbol of $\mathfrak{L}_M$, nor does $\mathfrak{L}_M$ have a comma. The comma and the double-barred turnstile belong to the **meta-language** we use to talk about the **object language** $\mathfrak{L}_M$. (The rest of our meta-language is mostly English.) The turnstile is used to express a certain relationship between $\mathfrak{L}_M$-sentence; it is not part of any $\mathfrak{L}_M$-sentence.

Clearly, if $\Diamond r$ and $\Box(r \to w)$ logically entail $\Diamond w$, then this doesn't depend on the choice of the letters r and w. We also have, for example:

$$\Diamond p, \Box(p \to q) \models \Diamond q$$

In general, it is part of the meaning of the turnstile (of 'logical consequence' or 'logical entailment') that uniformly replacing sentence letters by other sentence letters does not change facts about logical entailment.

We can generalize even further. Arguably, if the inference from $\Diamond p$ and $\Box(p \to q)$ to $\Diamond q$ is logically valid, then it remains valid if we replace the sentence letters p and q by arbitrary sentences, not just by other sentence letters. For example, if it might be cold and windy, and there is no doubt that the picnic does not take place if it is cold and windy, we can infer that it might be that the picnic does not take place:

$$\Diamond(c \wedge w), \Box((c \wedge w) \to \neg p) \models \Diamond \neg p$$

We can summarize the general pattern by a **schema**:

$$\Diamond A, \Box(A \to B) \models \Diamond B$$

Here 'A' and 'B' are placeholders for arbitrary $\mathfrak{L}_M$-sentences. The schematic statement means that if you plug in any $\mathfrak{L}_M$-sentences for A and B, the sentences to the left of the turnstile logically entail the sentence on the right. Anything that results from a schema by uniformly replacing the placeholders 'A', 'B', 'C', etc. with object-language sentences is called an **instance** of the schema. "Uniformly" means that the same schematic letter ('A', 'B', 'C', etc.) is always replaced by the same object-language sentence. (It is not required that different schematic letters are replaced by different object-language sentence.)

Exercise 1.3

Which of the following (meta-language) statements are instances of $\Diamond A, \Box(A \to B) \models \Diamond B$?

(a) $\Diamond p, \Box(p \to (q \wedge r)) \models \Diamond(q \wedge r)$
(b) $\Diamond p, \Box(p \to p) \models \Diamond p$
(c) $\Diamond(A \wedge B), \Box((A \wedge B) \to C) \models \Diamond C$
(d) $\Diamond\Box r, \Box(\Diamond\Box r \to \neg(p \wedge q)) \models \Diamond\neg(p \wedge q)$
(e) $\Diamond\Diamond p, \Box(\Diamond p \to \neg\Box(q \wedge r)) \models \Diamond\neg\Box(q \wedge r)$

So far, we have met one inference pattern that appears to be valid on the interpretation we have given to the box and the diamond: from $\Diamond A$ and $\Box(A \to B)$ one can infer $\Diamond B$. Let's look at other such patterns.

We can plausibly assume that modal propositional logic inherits the valid inference patterns from non-modal propositional logic. For example, any instance of $A \wedge B$ plausibly entails the relevant instance of A, even if the instances contain modal operators. This makes our modal logic an **extension** of classical propositional logic.

A useful fact about entailment in classical logic, which carries over to extensions of classical logic, is this:

Observation 1.1: If Γ ('gamma') is a list of sentences and A and B are sentences, then

$$\Gamma, A \models B \text{ iff } \Gamma \models A \to B$$

Proof. For a rigorous proof of observation 1.1, we would need a precise definition of the turnstile. (I will give such a definition in chapter 2.) But the intuitive reason is easy enough to understand. Let me explain the left-to-right direction, that if $\Gamma, A \models B$ then $\Gamma \models A \to B$. The right-to-left direction is similar.

The argument is by contraposition. I will show that if $\Gamma \models A \to B$ is *not* the case, then $\Gamma, A \models B$ is not the case either. So assume that for some sentences A, B, and some list Γ, it is not the case that $\Gamma \models A \to B$. This means that there is a conceivable scenario in which the sentences in Γ are all true while $A \to B$ is false, on some interpretation of the sentence letters. By the truth table for the material conditional, $A \to B$ is false only if A is true and B is false. So in the relevant scenario, the

sentences in Γ and A are true and B is false (under the given interpretation of the sentence letters). And then $\Gamma, A \models B$ is false, for $\Gamma, A \models B$ states that there is no conceivable scenario in which the sentences in Γ and A are all true while B is false, under any interpretation of the sentence letters. □

Observation 1.1 tells us that we can always move the turnstile to the left and put an arrow in its original position. For example, instead of

$$\Box(p \to q), \Diamond p \models \Diamond q$$

we can equivalently say

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

We can even go one step further to

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

This says that $\Box(p \to q) \to (\Diamond p \to \Diamond q)$ logically follows from no premises at all: the sentence is true in all conceivable scenarios under all interpretations of the sentence letters. Sentences like this are called **logically true** or **(logically) valid**.

(So an *argument* is called valid if the conclusion follows from the premises, while a *sentence* is called valid if it follows from no premises.)

Make sure you don't confuse the arrow with the turnstile. For one, the two symbols belong to different languages. The arrow is part of the object-language $\mathfrak{L}_M$, while the turnstile is part of our meta-language. Moreover, the two symbols have very different meanings. In $\mathfrak{L}_M$, $p \to q$ is true iff either p is false or q is true. By contrast, $p \models q$ is true iff there is no conceivable scenario in which p is true and q is false, on any interpretation of the two letters. Nonetheless, there is an important connection between the arrow and the turnstile: $A \models B$ is *true* iff $A \to B$ is *valid*. This connection is generalised by observation 1.1.

A practical upshot of observation 1.1 is that instead of asking which sentences in $\mathfrak{L}_M$ entail which other sentences, we can equivalently ask which sentences are valid. This is how the question is often framed in modal logic. For example, our earlier claim that $\Box(A \to B), \Diamond A \models \Diamond B$ can be re-stated as the claim that all instances of the following schema are valid:

$$\Box(A \to B) \to (\Diamond A \to \Diamond B) \qquad (\mathbf{K}^*)$$

If all instances of a schema are valid, we say that the schema itself is valid. (Note that $\mathbf{K}^*$ no longer contains a turnstile; the instances of the schema are simply $\mathfrak{L}_M$-sentences.)

1.3 Some modal schemas

The schema $\mathbf{K}^*$ is closely related to a more famous schema known as **K** (after Saul Kripke):

$$\Box(A \to B) \to (\Box A \to \Box B) \qquad (\mathbf{K})$$

On the present reading of the box as 'it is certain that', treating **K** as valid is to assume that if $A \to B$ and A are both certain, then it logically follows that B is also certain.

I will explain how **K** and $\mathbf{K}^*$ are related in a moment. First, I want to introduce two other schemas, connecting the box and the diamond.

$$\neg\Box A \leftrightarrow \Diamond\neg A \qquad (\mathbf{Dual1})$$

$$\neg\Diamond A \leftrightarrow \Box\neg A \qquad (\mathbf{Dual2})$$

Both of these are plausibly valid on our interpretation of the box and the diamond. Take **Dual1**. From left to right, this says that if it is not certain that A is true, then A might be false. From right to left, it states that if A might be false, then it is not certain that A is true. Similarly, from left to right, **Dual2** states that if it is not the case that A might be true, then it is certain that A is false; from right to left, it states that if it is certain that A is false, then it is not the case that A might be true. I hope you agree that these claims sound plausible.

Dual1 and **Dual2** are equivalent (in classical propositional logic) to the following schemas, as you should check:

$$\Box A \leftrightarrow \neg\Diamond\neg A \qquad (\mathbf{Dual1})$$

$$\Diamond A \leftrightarrow \neg\Box\neg A \qquad (\mathbf{Dual2})$$

So, if we wanted, we could define the box in terms of the diamond, or the diamond in terms of the box.

If two operators stand in the relationship expressed by **Dual1** and **Dual2**, they are called **duals** of each other. 'It might be that' and 'it is certain that' are plausibly

duals, but so are many other pairs of expressions in natural language. In modal logic, there is a convention to use the symbols $\Box$ and $\Diamond$ only for concepts that are duals of each other.

Exercise 1.4

Find all pairs of duals among the following English expressions.

(a) It is necessary that . . .
(b) It is impossible that . . .
(c) It is possible that . . .
(d) It is possibly not the case that . . .
(e) It was the case that . . .
(f) It will be the case that . . .
(g) It has always been the case that . . .
(h) It will always be the case that . . .
(i) The law requires that . . .
(j) The law does not require that . . .
(k) The law allows that . . .
(l) It is true that . . .
(m) It is false that . . .

Now I can explain how $\mathbf{K}^*$ is related to $\mathbf{K}$: the two schemas are plausibly equivalent, in the sense that the validity of one entails the validity of the other. Again, I'll only show one direction, because the other direction is similar. I'll show that if $\mathbf{K}$ is valid, then so is $\mathbf{K}^*$.

So suppose that $\mathbf{K}$ is valid; that is, every instance of $\mathbf{K}$ is a logic truth. Then so is every instance of

$$\Box(\neg A \to \neg B) \to (\Box\neg A \to \Box\neg B)$$

because every instance of this schema *is* an instance of $\mathbf{K}$.

To proceed, I need a further assumption. I'll assume that if two sentences are logically equivalent, then replacing one by the other in a more complex sentence does not affect whether the larger sentence is logically valid. On that assumption, we

can replace $\neg A \to \neg B$ in the previous schema by $B \to A$:

$$\Box(B \to A) \to (\Box\neg A \to \Box\neg B)$$

Moreover, assuming that **Dual2** is valid, $\Box\neg A \to \Box\neg B$ is logically equivalent to $\neg\Diamond A \to \neg\Diamond B$, which in turn is equivalent to $\Diamond B \to \Diamond A$. So we get

$$\Box(B \to A) \to (\Diamond B \to \Diamond A),$$

which is obviously equivalent to $\mathbf{K}^*$.

Exercise 1.5

Spell out the converse argument, that if $\mathbf{K}^*$ is valid, then so is **K**.

Here are some other famous principles from modal logic; these are not entailed by **K**.

$$\Box A \to A \qquad (\mathbf{T})$$
$$\Box A \to \Diamond A \qquad (\mathbf{D})$$
$$\Box A \to \Box\Box A \qquad (\mathbf{4})$$
$$\Diamond A \to \Box\Diamond A \qquad (\mathbf{5})$$

A few comments on the names. The first schema is called **T** because its validity means that $\Box A$ implies the *truth* of A. The second is called **D** because it plays an important role in a branch of modal logic called *deontic logic*, where the box is read as 'it is obligatory that' and the diamond as 'it is permitted that'. The labels **4** and **5** allude to a list of logical "systems" discussed by C.I. Lewis in the late 1920s and early 1930s. **4** is the characteristic principle of the fourth system ('S4') in Lewis's list; **5** is the characteristic principle of the fifth system ('S5').

Which of these schemas should we accept as valid if we read the box as 'it is certain that' and the diamond as 'it might be that'?

Consider **T**. If it is certain that A, does it logically follow that A is true? Maybe not. What do you think?

The validity of **D** would mean that whenever it is certain that A, then A might be the case. Now, if it is certain that A, it would be odd to say that A *might* be the case.

But that something sounds odd doesn't mean that it is false. If we are certain that A, and the question arises whether A might be true, the correct answer is arguably 'yes', not 'no'. So I'd say that **D** is valid.

What about **4** and **5**? If something is certain, can we infer that it is certain that it is certain? If something might be the case, can we infer that it is certain that it might be the case? Again, the answers aren't obvious.

It gets worse. Consider the following schema, named after Peter Geach.

$$\Diamond\Box A \rightarrow \Box\Diamond A \qquad \textbf{(G)}$$

If it might be the case that something is certain, is it certain that it might be the case? What does that even mean?

The kind of problem we here encounter comes up often in modal logic. We start with an intuitive concept, such as the concept that something might be the case. We represent this concept by a new symbol in a formal language. When we then consider which inferences involving the new symbol should count as valid, we realize that the intuitive concept with which we began does not give a clear answer. We need to sharpen the concept, giving it a clearer meaning, if we want to define its logic.

Exercise 1.6

Explain why, if **T** is valid, then so is the converse of **4**, $\Box\Box A \rightarrow \Box A$.

Exercise 1.7

Show that **T** is valid iff $A \rightarrow \Diamond A$ is valid, assuming the validity of **Dual1** and **Dual2**.

1.4 Flavours of modality

We have looked at one application of modal logic, in which we introduced sentence operators for 'it might be' and 'it is certain'. There are many other applications.

In philosophy and linguistics, a statement is classified as *modal* if it is about what *must* or *may* or *might* or *can* or *could have* been the case, in some sense of 'must', 'may', 'might', 'can', or 'could have'. So the following statements would be classified as modal.

(1) It must be raining.
(2) We might get wet.
(3) There can't be any water left.
(4) It can take years to earn someone's trust.
(5) It could have been raining.
(6) You can't go from Auckland to Sydney by train.
(7) In chess, you must be ruthless.
(8) You may leave now.
(9) You can take the 41 bus to get to the railway station.

Modal statements don't have to involve an auxiliary verb like 'might', 'must', or 'may'. We can, for example, use adjectives like 'possible' and 'necessary' to talk about what might or must be the case. Or we can use adverbs like 'possibly' and 'necessarily'. Verbs like 'have to' and 'ought to' also have a modal meaning, as do suffixes like '-ble' in 'comprehensible', 'legible', or 'edible'.

So the class of modal statements is grammatically diverse. In terms of meaning, we can distinguish at least three groups, often called *flavours* of modality.

The first group are statements about what is known, or entailed by the available evidence. Examples (1)–(3) fall into this category, at least on their most natural usage. This flavour of modality is called *epistemic* (from Greek *episteme*: 'knowledge').

A second flavour of modality, illustrated by examples (7)–(9), is concerned with rules, prescriptions, norms, or with what is required to achieve a certain goal. This flavour of modality is called *deontic* (from Greek *deontos*: 'of that which is binding').

Examples (4)–(6) would normally be understood as neither epistemic nor deontic. When I say that you can't go from Auckland to Sydney by train, I don't just mean that my information implies that you *don't* go from Auckland to Sydney by train; nor do I mean that you're not permitted to go, by some relevant norms. Rather, I mean that relevant circumstances in the world – such as the presence of an ocean between Auckland and Sydney – make it impossible for you to travel the journey by train. This flavour of modality is sometimes called *circumstantial*. It comes in many sub-flavours, depending on what kinds of circumstances are considered.

Each of these flavours of modality corresponds to a branch of modal logic.

Epistemic logic is a branch of modal logic that formalizes reasoning about knowledge and information. When we understood the diamond in $\mathfrak{Q}_M$ as 'it might be

that' and the box as 'it is certain that', we were doing epistemic logic. We will return to epistemic logic in chapter 5.

Deontic logic is another branch of modal logic, concerned with norms, permissions, and obligations. Standard deontic logic has a box-like sentence operator for 'it is obligatory that' (or 'it must be that', in the deontic sense), and a diamond-like operator for 'it is permissible that' (or 'it may be that', in the deontic sense). We will look at deontic logic in chapter 6.

A third branch of modal logic might be called *circumstantial logic*, but nobody uses that label. Some authors speak of *alethic modal logic* (from *aletheia*: 'truth'), but that label is also not used widely, and it is used for different things by different authors.

Two sub-flavours of circumstantial modality deserve special mention, because much philosophical work in modal logic concentrates on these sub-flavours.

The first sub-flavour is known as *metaphysical* modality. Metaphysical modality is concerned with what is compatible or incompatible with the nature of things. For example, many philosophers have the intuition that part of what it is to be water – part of the very nature of water – is to contain hydrogen. In philosophy jargon, this means that it is "metaphysically necessary" that water contains hydrogen. To formalize reasoning about metaphysical modality, we can introduce a box-like operator for metaphysical necessity, and a diamond-like operator for the dual concept of metaphysical possibility. The term 'alethic modal logic' is sometimes reserved for this sub-branch of circumstantial modal logic.

The other sub-flavour is variously known as *logical* or *absolute* or *unrestricted* modality. Here, the guiding intuition is that when we consider whether something is circumstantially possible – say, whether one could travel from Auckland to Sydney by train – we normally ignore various possibilities that are incompatible with conversationally relevant facts or circumstances. For example, we almost always ignore possible scenarios in which the laws of nature or the geography of the Earth are different. If there were no ocean between Auckland and Sydney, and someone had put a railway line there, one certainly could travel from Auckland to Sydney by train. So it isn't *absolutely* or *logically* impossible to travel from Auckland to Sydney by train. Nor is it absolutely impossible to travel the 2,200 km route in 1 millisecond, even though that contradicts our laws of physics. Something is absolutely (logically, unrestrictedly) impossible only if there is no way it could have been the case, not ignoring any way things could have been. For example, it is absolutely impossible to

travel the 2,200 km from Auckland to Sydney in 1 millisecond while travelling at an average speed of 100 km/h.

Some philosophers hold that metaphysical modality and absolute modality coincide: something is absolutely possible/impossible just in case it is metaphysically possible/impossible. Others hold that metaphysical possibility is more restricted than absolute possibility. Yet others hold that there is no such thing as metaphysical modality, or no such thing as absolute modality. We won't enter into these debates.

From a logical perspective, the idea of absolute modality is attractive because the logic of absolute modality is especially simple and well-behaved. As we will see in the next chapter, if we interpret the box and the diamond as absolute necessity and possibility, then the schemas **T**, **D**, **K**, **4**, **5**, and **G** all plausibly become valid, and there is a simple method for checking whether any schema, no matter how complex, is valid.

Exercise 1.8

Which of **T**, **K**, and **4** do you think are valid if we read the box as 'it is obligatory that' and the diamond as 'it is permitted that'?

Exercise 1.9

Translate the following sentences, as well as possible, into $\mathfrak{L}_M$, assuming that the diamond expresses epistemic possibility ('it might be that') and the box epistemic necessity ('it must be that').

(a) I may have offended the principal.
(b) It can't be raining.
(c) Perhaps there is life on Mars.
(d) If the murderer escaped through the window, there must be traces on the ground.

Exercise 1.10

Translate the following sentences, as well as possible, into $\mathfrak{L}_M$, assuming that the diamond expresses deontic possibility ('it is permitted that') and the box deontic necessity ('it is obligatory that').

(a) I must go home.
(b) You don't have to come.
(c) You can't have another beer.
(d) If you don't have a ticket, you must pay a fine.
(e) You need a special visa to enter Chukotka.

Exercise 1.11

Translate the following sentences, as well as possible, into $\mathfrak{L}_M$, assuming that the diamond expresses (some relevant sub-flavour of) circumstantial possibility and the box circumstantial necessity.

(a) It could have snowed today.
(b) It's impossible for me to both cook and entertain the children.
(c) I can't hear you if you're talking to me from the kitchen.
(d) If you can't go to the station, you can't take the train.

1.5 Beyond modality

Despite its name, modal logic extends well beyond the study of modality.

In chapter 7 we will look at a branch of modal logic used to reason about the flow of time, called *temporal logic*. Here we will have operators that function like 'it was the case that' and 'it will be the case that'.

In chapter 8, we will turn to *conditional logic*. Here we will introduce several (non-truth-functional) two-place operators and investigate to what extent they match certain 'if . . . then . . . ' constructions in English.

In chapter 4, we will briefly look at *provability logic*, which studies formal properties of mathematical provability. Here the box is read as 'it is mathematically provable that'.

There are many other branches of modal logic that we won't be able to cover. For example, one lively branch is *dynamic logic*, which is used to reason about how certain actions or events change the state of some system. For any relevant action x, dynamic logic introduces a box-like operator $[x]$ and a diamond-like operator $\langle x\rangle$; $[x]A$ means that the action x will definitely lead to outcome A, while $\langle x\rangle A$ means

that x may lead to A. Dynamic logic is widely used in computer science, where the relevant "actions" are typically steps of a computer program.

The many branches of modal logic aren't isolated disciplines. There is a reason why they are commonly treated under the unifying heading of modal logic. Many tools and techniques that have been developed for one branch can also be used for others, and many of the applications share a common abstract core.

What's more, the different branches are often combined. For example, *dynamic epistemic logic* combines ideas from dynamic and epistemic logic to model how knowledge and information change across time. (We will covertly do a bit of dynamic epistemic logic in chapter 5.) It can also be useful to combine, say, deontic and epistemic logic, to reason about what people know about their obligations, or deontic and alethic/circumstancial logic, perhaps to scrutinize the idea that 'ought' implies 'can'.

To conclude this introductory chapter, let's have a quick look at a fun little application that connects to a topic we have discussed.

Suppose we add to the language of propositional modal logic a sentence operator $\Box$ so that $\Box A$ means that A is logically true – true in virtue of its logical form. On this interpretation, $\Box(p \vee \neg p)$ is true, because $p \vee \neg p$ is true in virtue of its form; $\Box(p \vee q)$ is false, because $p \vee q$ is not true in virtue of its form.

Let's revisit the schemas we have considered in section 1.2, beginning with **K**.

$$\Box(A \to B) \to (\Box A \to \Box B) \qquad (\mathbf{K})$$

If a conditional $A \to B$ is true in virtue of its form, and so is A, can we conclude that B is true in virtue of its form? Arguably yes. Intuitively, if A is true in any conceivable scenario under any interpretation of the sentence letters, and $A \to B$ is true in any conceivable scenario under any interpretation of the sentence letters, then so is B – for B is bound to be true in any scenario in which A and $A \to B$ are both true.

T is easier.

$$\Box A \to A \qquad (\mathbf{T})$$

If something is logically true, then we can surely infer (as a matter of logic) that it is true.

What about **4**?

$$\Box A \to \Box\Box A \qquad (\mathbf{4})$$

Take an example. $p \vee \neg p$ is logically true; so $\Box(p \vee \neg p)$ is true. You don't need to know what p means in order to see that $\Box(p \vee \neg p)$ is true, nor do you need to know any substantive facts about the world: the statement is true in virtue of its logical form. So $\Box\Box(p \vee \neg p)$ is true as well. In general, if A is true in virtue of its form, then $\Box A$ is also true in virtue of its form. So $\Box A$ does entail $\Box\Box A$.

Next, **5**:

$$\Diamond A \rightarrow \Box\Diamond A \tag{5}$$

I haven't introduced a diamond symbol yet; let's stipulate that $\Diamond A$ is the dual of $\Box A$. So $\Diamond A$ means that $\neg A$ is *not* true in virtue of its logical form. **Dual1** and **Dual2** are then trivially valid. **5** says that if $\neg A$ is not true in virtue of its form, then it is true in virtue its form that $\neg A$ is not true in virtue of its form. This isn't easy to understand, but the following line of thought shows that it is plausible.

Suppose the antecedent of **5** is true: $\neg A$ is not true in virtue of its form. If a sentence is not true in virtue of its form, then evidently any sentence of the same form also isn't true in virtue of its form. So, given that $\neg A$ is not true in virtue of its form, one can tell merely by the logical form of $\neg A$ that $\neg A$ is not true in virtue of its form – in other words, that $\Box\neg A$ is false. So one can tell merely by the logical form of $\Box\neg A$ that it is false. And so one can tell merely by the logical form of $\neg\Box\neg A$ – equivalently, $\Diamond A$ – that it is true. So from the assumption $\Diamond A$ we can logically infer $\Box\Diamond A$. So **5** is valid.

Exercise 1.12

Explain why **D** and **G** are valid on the present interpretation of the box and the diamond.

Exercise 1.13

(a) Which of the schemas we have considered are valid if the box is interpreted as 'it is true that' (and the diamond as the dual of the box)?

(b) Which of the schemas are valid if the box is interpreted as 'it is either true or false that' (and the diamond as the dual)?

2 Possible Worlds

2.1 The possible-worlds analysis of possibility and necessity

An important breakthrough in the history of modal logic was the development of "possible-worlds semantics" in the 1950s. The basic idea of possible-worlds semantics is to analyze possibility and necessity in terms of truth at a possible world:

> A proposition is possible iff it is true at some possible world.
> A proposition is necessary iff it is true at all possible worlds.

In philosophy, a **possible world** is commonly understood as a complete way things could have been. By comparison, here is an incomplete way things could have been: I could have had coffee for breakfast today. (In fact I had tea.) This is an incomplete way things could have been, because it leaves many things open: how much coffee I had, which cup I used, how fast I consumed the coffee, and so on. A possible world, by contrast, leaves nothing open; it settles every question.

An example of a possible world is the **actual world** – the totality of everything that is the case in our universe. There are many things we don't know about the actual world. We don't know whether there is life on other planets, we don't know who killed Richard Montague; but these questions have an answer. Every precise and unambiguous hypothesis about the actual world is either true or false. That's why the actual world is a *complete* way things could have been.

Now suppose we'd like to reason about what could or could not have been the case. To formalize this kind of reasoning, we can use the standard language $\mathfrak{L}_M$ of modal propositional logic, reading the diamond as 'it could have been the case that'. By convention, the box will be the dual of the diamond, so $\Box p$ will mean that it could *not* have been the case that *not-p*.

(Philosophers often use 'it is possible' to mean 'it could have been the case that',

and 'it is necessary' to mean 'it could not have failed to be the case that'. On that usage, it is possible that I had coffee for breakfast, even though I am sure that I had tea.)

Which argument forms or schemas should we count as valid on the present understanding of the box and the diamond? Does $\Box A$ entail $\Box\Box A$? Does $\Diamond\Box A$ entail $\Box\Diamond A$? If it could have been the case that it could not have been the case that not-A, does it follow that it could not have been the case that it could not have been the case that A? These questions are hard to answer directly. Possible-worlds semantics will clear things up.

The possible-worlds analysis of the box and the diamond effectively provides a translation from the modal language $\mathfrak{L}_M$ into a non-modal (meta-)language in which we don't have modal operators but instead quantify over possible worlds, talking about what is the case at *all* worlds, or at *some* worlds. These non-modal, quantificational claims are often easier to understand than the original modal claims.

The analysis can also been understood as a metaphysical reduction. On that view, modal facts about what could or could not have been the case are really, fundamentally, facts about what *is* the case at other possible worlds. Consider the fact that I could have had coffee this morning. Is this a primitive, inexplicable truth, or can it be explained by other facts? And if it can be explained by other facts, must these facts themselves be modal, or is there an explanation that only appeals to non-modal facts? Many philosophers hold that all modal truths are ultimately reducible to non-modal truths. And some of these philosophers have suggested that the possible-worlds analysis might show how the reduction works: modal truths about what could or could not have been the case are made true by non-modal facts about what is the case at other possible worlds – assuming one can give a non-modal analysis of 'possible world'. (This project is mainly associated with the philosopher David Lewis.)

Our topic is the logic, not the metaphysics of modality. So we won't take a stance on whether, or how, modal facts are reducible to non-modal facts. Almost everyone agrees that one can clarify questions about possibility and necessity by thinking in terms of possible worlds. That's all we need.

Indeed, if we take seriously the conception of a possible world as a complete way things could have been, then the hypothesis that something could have been the case iff it is the case at some possible world does not amount to much more than the claim that any incomplete way things could have been (like, me having coffee for breakfast) can be extended to a complete way things could have been. In other words, if p could

have been the case, then there is a way the entire world could have been that would have included p.

In practice, we don't even take the completeness of possible worlds all that seriously. We will often work with toy worlds that merely settle all the questions in which we're interested, leaving lots of other questions open.

2.2 Models

Before I spell out the possible-worlds semantics for $\mathfrak{L}_M$, I need to review what logicians mean by a "model".

Recall that an argument is *valid* if there is no conceivable scenario in which the premises are true and the conclusion is false; an argument is *logically valid* if it is valid merely in virtue of its logical form, so that there is no interpretation of the argument's descriptive vocabulary that makes the premises true and the conclusion false, in any conceivable scenario.

So logical validity is a matter of truth *in all conceivable scenarios under all interpretations of the descriptive vocabulary*. In logic, a *model* is basically a package of a scenario and an interpretation of descriptive vocabulary, so that validity can be defined in terms of truth at all models.

But logicians are economical, so they trim down their models to the smallest core that will do the job. Take classical propositional logic, without modal operators. If we want to figure out whether a sentence is true or false in a particular scenario under some interpretation of the sentence letters, all we really need to know is which sentence letters are true and which are false in the scenario, under the given interpretation of the sentence letters. That's because all operators in classical propositional logic are truth-functional. So we can ignore everything about the scenario and the interpretation of the sentence letters except which sentence letters come out true and which come out false. Hence logicians identify a model of classical propositional logic simply with an assignment of truth-values to the sentence letters. You can specify such a model by saying, for example, that p is true and all other sentence letters are false. Intuitively, that isn't much of a scenario, nor have you thereby explained what the sentence letters mean, in any intuitive sense. But it does the job.

You may wonder why we need to consider alternative scenarios *and* alternative interpretations of the descriptive vocabulary to define validity. Can't we say that

an argument is valid iff any interpretation of the descriptive vocabulary that makes the premises true also makes the conclusion true, without looking at alternative scenarios?

That would work for classical propositional logic, but not for more expressive languages. Consider the sentence $\exists x \exists y \neg(x = y)$ in the language of predicate logic. If we count the identity symbol as logical, this sentence contains no descriptive terms at all. And the sentence is true, because there is in fact more than one object. So the sentence is true under any interpretation of its non-logical vocabulary. But it shouldn't count as *logically* true; it doesn't logically follow from any premises whatsoever. The sentence is false in any scenario in which there is only one object.

Unlike models of propositional logics, a model of predicate logic therefore doesn't just give a (partial) interpretation of the non-logical expressions, but also specifies a domain of individuals over which the quantifiers are taken to range. $\exists x \exists y \neg(x = y)$ is false in any model whose individual domain contains only a single object.

Now, what do we need from a model in modal propositional logic? We know that it is not enough to fix the truth-values of the sentence letters, if the box and the diamond are supposed to represent any form of necessity and possibility, since these notions aren't truth-functional. Following the possible-worlds analysis, we will instead assume that a model fixes the truth-value of each sentence letter at each possible world. We can then figure out whether, for example, $\Box p$ is true in a model by checking whether p is true at all worlds.

How should we represent the possible worlds? In the early days of possible-worlds semantics, philosophers thought that a possible world is more or less the same thing as a conceivable scenario, and they often identified the space of possible worlds with the class of all models of predicate logic. However, it has proved useful to treat the space of possible worlds as a non-logical matter, so that different models may involve different possible worlds, just as different models of predicate logic may involve different sets of individuals.

So we will define a model of modal propositional logic as consisting of, first, a set of worlds, and second, an interpretation function that assigns a truth-value to each sentence letter at each world. Again, this should be understood as a trimmed-down package of a conceivable scenario and an interpretation of the descriptive vocabulary, stripped of any information that isn't necessary to settle facts about validity in modal propositional logic.

Definition 2.1

A **(basic) model** of $\mathfrak{L}_M$ is a pair $\langle W, V \rangle$ consisting of

- a non-empty set W, and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value (*True*=1 or *False*=0).

In the next chapter, we will replace this definition by a slightly more complicated definition, which is why I've called models of the present kind 'basic'.

The interpretation V (for 'valuation') in a model takes two arguments as input: a sentence letter and a world. You can picture an interpretation function as a table:

	w_1	w_2	w_3	w_4	...
p	1	1	0	1	
q	0	0	1	1	
r	1	0	0	0	
$\vdots$					

Note that V only fixes the truth-values of sentence letters. The truth-value of more complex sentences is then determined by the meaning of the logical operators. For example, if p and q are both true at some world w, then the meaning of $\wedge$ ensures that $p \wedge q$ is also true at w.

To define the interpretation of complex sentences at a world, and thereby the meaning of the logical operators, I will use (meta-linguistic) statements of the form

$$M, w \models A$$

as shorthand for

$$A \text{ is true at world } w \text{ in model } M.$$

I will use '$M, w \not\models A$' as the negation of '$M, w \models A$'.

Yes, it's the same double-barred turnstile that we also use for logical consequence. This should cause no confusion because it is usually clear if the things to the left of the turnstile are $\mathfrak{L}_M$-sentences or meta-linguistic expressions for a model and a world.

Formally, the relation $\models$ between a model, a world and an $\mathfrak{L}_M$-sentence is defined as follows.

Definition 2.2: Basic Possible-Worlds Semantics

If $M = \langle W, V \rangle$ is a basic model, w is a member of W, ρ is any sentence letter, and A, B are any $\mathfrak{L}_M$-sentences, then

(a)	$M, w \models \rho$	iff $V(\rho, w) = 1$.
(b)	$M, w \models \neg A$	iff $M, w \not\models A$.
(c)	$M, w \models A \wedge B$	iff $M, w \models A$ and $M, w \models B$.
(d)	$M, w \models A \vee B$	iff $M, w \models A$ or $M, w \models B$.
(e)	$M, w \models A \rightarrow B$	iff $M, w \models B$ or $M, w \not\models A$.
(f)	$M, w \models A \leftrightarrow B$	iff $M, w \models (A \rightarrow B)$ and $M, w \models (B \rightarrow A)$.
(g)	$M, w \models \Box A$	iff $M, v \models A$ for all v in W.
(h)	$M, w \models \Diamond A$	iff $M, v \models A$ for some v in W.

Let me explain. Any model M contains an interpretation function V. Intuitively, V tells us which sentence letters are true and which are false at any world in the model. Clause (a) makes this explicit. It says that a sentence letter ρ is true at a world w in a model $\langle W, V \rangle$ iff V assigns *True* (=1) to ρ and w.

Clause (b) tells us that the negation $\neg A$ of an $\mathfrak{L}_M$-sentence A is true at a world in a model iff A is not true at that world in that model. This basically means that the truth-table for negation applies locally at every world: at any world, $\neg A$ is true iff A is not true.

Clauses (c)–(f) similarly tell us that the truth-tables for the other truth-functional connectives apply locally at each world. For example, (c) says that the conjunction $A \wedge B$ of two $\mathfrak{L}_M$-sentences A and B is true at a world in a model iff both A and B are true at that world in that model.

Finally, clauses (g) and (h) spell out the possible-worlds analysis of the box and the diamond. According to (g), a sentence $\Box A$ is true at a world in a model iff A is true at all worlds in the model. According to (h), $\Diamond A$ is true at a world in a model iff A is true at some world in the same model.

The whole definition is called a *semantics* because a semantics for a language is an account of what the expressions in the language mean, and definition 2.2 can be seen as giving the meaning of the logical expressions in $\mathfrak{L}_M$.

Definition 2.2 settles the truth-value of every sentence at every world in every model, because every $\mathfrak{L}_M$-sentence is built up from sentence letters with the operators

covered in definition 2.2.

To illustrate, consider the following partial specification of a model M:

$$W = \{w, v\}$$
$$V(p, w) = 1, V(p, v) = 1$$
$$V(q, w) = 1, V(q, v) = 0$$

This model contains only two worlds, w and v; the interpretation function V makes p true at both worlds, and it makes q true at w but not v. I have left the other proposition letters uninterpreted. With the help of definition 2.2, we can figure out at which of the two worlds, say, $\Box\Diamond(\Box q \rightarrow \Diamond\Box p)$ is true. We start with the smallest parts of the sentence.

1. p is true at w and v (by clause (a) of definition 2.2).
2. q is true at w and not true at v (by clause (a) of definition 2.2).
3. $\Box p$ is true at w and v (by 1 and clause (g) definition 2.2).
4. $\Box q$ is true at no world (by 2 and clause (g) of definition 2.2).
5. $\Diamond\Box p$ is true at w and v (by 3 and clause (h) of definition 2.2).
6. $(\Box q \rightarrow \Diamond\Box p)$ is true at w and v (by 4, 5, and clause (e) of definition 2.2).
7. $\Diamond(\Box p \rightarrow \Diamond\Box q)$ is true at w and v (by 6 and clause (h) of definition 2.2).
8. $\Box\Diamond(\Diamond p \rightarrow \Diamond\Box q)$ is true at w and v (by 7 and clause (g) of definition 2.2).

Exercise 2.1

At which worlds in the model just described is $\Diamond p \rightarrow (q \vee \Diamond\Box p)$ true?

2.3 Validity

A somewhat odd feature of our semantics is that $\mathfrak{L}_M$-sentences are true or false only relative to a model *and a world*. In this respect, modal logic differs from classical propositional and predicate logic. In classical logic, sentences are true or false relative to a model. In modal logic, you also need to know at which point within a model – at which world – a sentence should be evaluated.

So we can't quite define validity or logical consequence in terms of truth at a model. Instead, we'll say that an argument is valid if it preserves truth at all worlds in all models:

Definition 2.3

A sentence A is a **logical consequence** of a set Γ of sentences (for short: $\Gamma \models A$) iff A is true at every world in every (basic) model at which all members of Γ are true.

Equivalently: A is a logical consequence of Γ if there is no world in any model at which all members of Γ are true while A is false.

Recall that a sentence is logically true (or valid) if it logically follows from no premises at all. By definition 2.3, this is equivalent to the following:

Definition 2.4

A sentence A is **valid** (for short: $\models A$) iff it is true at every world in every model.

As in the previous chapter, we will also apply the concepts of validity and logical consequence to sentence schemas. For example, the schematic statement

$$A \wedge B \models A$$

expresses that every instance of the schema $A \wedge B$ logically entails the corresponding instance of A. Similarly, the schematic

$$\models (A \wedge B) \rightarrow A$$

means that every instance of the schema $(A \wedge B) \rightarrow A$ is valid.

Now that we have a formal definition of the turnstile, we can verify some claims I made in the previous chapter.

To begin, we could give a rigorous proof of observation 1.1 from p.11, that $\Gamma, A \models B$ iff $\Gamma \models A \rightarrow B$. I will not go through the proof, however, because it would look much like the informal argument I gave in section 1.2.

I also claimed, on p.11, that modal propositional logic is an extension of classical propositional logic, in the following sense:

Propositional Extension Theorem

Whenever a sentence A is a truth-functional consequence of a set of sentences Γ, then $\Gamma \models A$.

A is a *truth-functional consequence* of Γ if the standard truth-tables for the propositional connectives guarantee that A is true whenever the sentences in Γ are all true. For example, the truth-table for $\rightarrow$ guarantees that if p and $p \rightarrow q$ are both true, then so is q; hence q is a truth-functional consequence of $\{p, p \rightarrow q\}$. By the propositional extension theorem, it follows that the inference from p and $p \rightarrow q$ to q is also valid in modal propositional logic; that is,

$$p, p \rightarrow q \models q.$$

Why is that? By definition 2.3, '$p, p \rightarrow q \models q$' means that whenever p and $p \rightarrow q$ are true at a world in a model, then so is q. Now look at definition 2.2. According to clause (e), $p \rightarrow q$ is true at a world iff p is false at that world or q is true. So if p and $p \rightarrow q$ are both true at a world, then q can't be false at that world.

In general, the modal extension theorem holds because the truth-functional connectives have their standard truth-table meaning relative to every world.

Lastly, on p.14, I made the following claim:

Replacement Theorem

If two sentences A and B are logically equivalent, then replacing one by the other within a more complex sentence C does not affect whether C is valid.

Given definition 2.4, this clearly follows from the following hypothesis, which I'll call '(*)':

(*) If two sentences A and B are logically equivalent, then replacing one by the other within a more complex sentence C does not affect whether C is true at any world w in any model M.

We can establish (*) by going through different possibilities about the logical form of C.

1. Suppose C is a sentence letter. In that case, there are no logically equivalent parts of C, so trivially, replacing any such parts by one another does not affect the truth-value of C at any world in any model.

2. Next, suppose that C is the negation of some other sentence D. We show that *if* (*) holds for D, *then* it also holds for C. So let C' be the result of replacing logically equivalent sentences within C, and let D' be the result of making the same replacements in D. Note that D' is $\neg D$. We have to show that if C and C' have the same truth-value at a world, then so do D and D'. This follows from clause (b) of definition 2.2, which ensures that at any world, D and D' have the opposite truth-value of C and C', respectively.

3. Next, suppose that C is the conjunction of two sentences D and E. Much like in the previous case, you can verify by clause (c) of definition 2.2 that if (*) holds for D and E, then (*) also holds for C.

I won't bore you by going through all the remaining cases: that C is disjunction $D \vee E$, a conditional $D \rightarrow E$, a biconditional $D \leftrightarrow E$, a box sentence $\Box D$, and a diamond sentence $\Diamond D$. In each case, it follows from definition 2.2 that if (*) holds for the immediate parts of C (for example, for D and E, if C is $D \vee E$), then (*) also holds for C. It then follows that (*) holds for all sentences whatsoever, because every $\mathfrak{L}_M$-sentence is built up from sentence letters by the operators $\neg$, $\wedge$, $\vee$, etc.

This style of argument is called an **induction on complexity** and is widely used when reasoning about formal languages. In general, if you want to show that every sentence in a formal language has some property, it suffices to show that (a) the smallest sentences in the language all have the property, and (b) *if* the immediate parts of a complex sentence have the property, then so does the complex sentence itself.

2.4 The logic of unrestricted modality

By definition 2.4, a sentence is valid iff it is true at all worlds in all models. Definition 2.1 tells us what counts as a model, and definition 2.2 determines the truth-value of any sentence at any world in any model. Together, these definitions therefore settle which sentences, and which schemas, are valid and which aren't.

Let's look at the **T** schema from the previous chapter.

$$\Box A \rightarrow A \qquad (\mathbf{T})$$

This turns out to be valid. To show this, we have to show that all instances of the schema are true at all worlds in all models. So let w be an arbitrary world in an arbitrary model M. Now, whatever $\mathfrak{L}_M$ sentence we plug in as A, either that sentence is true at w in M or not. If A is true at w in M, then by clause (e) of definition 2.2, $\Box A \rightarrow A$ is also true at w in M. If A is not true at w in M, then by clause (g) of definition 2.2, $\Box A$ is not true at w in M either, and then $\Box A \rightarrow A$ is true at w by clause (e). So either way, $\Box A \rightarrow A$ is true at w in M. Since w and M were chosen arbitrarily, this means that every instance of $\Box A \rightarrow A$ is true at every world in every model. So **T** is valid.

How about, say, schema **4**?

$$\Box A \rightarrow \Box\Box A \qquad (\mathbf{4})$$

If something is necessary, is it necessarily necessary? Our possible-worlds semantics says yes. As before, let w be an arbitrary world in an arbitrary model. If $\Box A$ is false at w, then $\Box A \rightarrow \Box\Box A$ is true at w, by clause (e) of definition 2.2. Suppose then that $\Box A$ is true at w. In that case, A is true at all worlds, by clause (g) of definition 2.2. And then $\Box A$ is true at all worlds, again by clause (g). And then, once again by clause (g), $\Box\Box A$ is true at all worlds. So whenever $\Box A$ is true at a world in a model, then so is $\Box\Box A$. By clause (e) of definition 2.2, it follows that $\Box A \rightarrow \Box\Box A$ is true at every world in every model.

We could continue with the other schemas from the previous chapter: $\mathbf{K}^*$, **K**, **Dual1**, **Dual2**, **D**, **5**, and **G**. As you can check, they all come out valid.

Exercise 2.2

Show that the **D**-schema $\Box A \rightarrow \Diamond A$ is valid.

You may have noticed, when working through definition 2.2, that if a sentence starts with a modal operator, then its truth-value no longer varies from world to world. Moreover, its truth-value doesn't change if you stack further modal operators in front of it. (This second observation actually follows from the first. Can you see why?) For example, if $\Diamond p$ is true at some world w in some model, then $\Diamond p$ is true at all

worlds in the model, and so $\Box\Diamond p$ is true at w as well, as is $\Diamond\Diamond p$.

It follows that on the present semantics, any sentence that begins with a sequence of modal operators is equivalent to the same sentence with all but the last operator removed. For example, $\Diamond\Box\Box\Diamond\Diamond p$ is equivalent to $\Diamond p$.

This is often useful to quickly check whether a schema is valid. For example, since any instance of $\Box\Box A$ is equivalent to the corresponding instance of $\Box A$, and we can always replace logically equivalent sentences within larger sentences, schema **4** is equivalent to $\Box A \rightarrow \Box A$. That's obviously valid. So we can see that **4** is valid, without going through the tedious argument above.

Exercise 2.3

Explain why **5** and **G** are valid, using the fact just mentioned.

Make sure you don't conflate the concepts of necessity and validity. Necessity means truth at all worlds (or so we currently assume). Validity means truth at all worlds *in all models*. Whether an $\mathfrak{L}_M$ sentence is necessary generally varies from model to model. In a model whose interpretation function assigns 1 to p relative to each world, p is necessary insofar as $\Box p$ is true at all worlds in the model. In other models, $\Box p$ is not true at all worlds. Validity, by contrast, is not relative to a model. The sentence p is definitely not valid. The sentence $\Box p \rightarrow p$ is.

Exercise 2.4

Show that if a sentence A is valid, then so is $\Box A$.

Here is an example of an invalid schema:

$$\Box(A \vee B) \rightarrow (\Box A \vee \Box B)$$

A schema is invalid if it has at least one instance that isn't valid. A relevant instance is

$$\Box(p \vee q) \rightarrow (\Box p \vee \Box q).$$

How could we show that this isn't valid?

By definition 2.4, a sentence is valid iff it is true at all worlds in all models. So we have to find some model in which there is some world at which the above sentence is

false. Such a model is called a **countermodel** for the sentence (or schema) we want to reveal as invalid.

There are many countermodels for $\Box(p \vee q) \rightarrow (\Box p \vee \Box q)$. Here is one, as you should verify with the help of definition 2.2.

$$W = \{w, v\}$$
$$V(p, w) = 1, V(p, v) = 0$$
$$V(q, w) = 0, V(q, v) = 1$$

Again this is not a complete model because I have ignored what V says about sentence letters other than p and q, which clearly wouldn't make a difference to $\Box(p \vee q) \rightarrow (\Box p \vee \Box q)$.

In general, to specify a countermodel for a sentence A, you have to specify two things: a set W of worlds, and an interpretation function V that assigns truth values to the sentences letters in A relative to each member of W.

Exercise 2.5

Show that the schema $A \rightarrow \Box A$ is invalid. (So we do not have $A \models \Box A$. Compare the previous exercise.)

Exercise 2.6

Show that if $\models A \rightarrow B$, then also $\models \Box A \rightarrow \Box B$.

2.5 Trees

Working through definition 2.2 to check a sentence (or schema) for validity is tiring and error-prone. I will now introduce a more elegant technique: the method of **semantic tableaux** or **trees**. (You may be familiar with the method for non-modal logic.) The method is in the first place a technique to find countermodels. It is best introduced by example.

Let's try to find a countermodel for $\Diamond p \rightarrow \Box p$. That is, we want to construct a model in which there is some world w at which $\Diamond p \rightarrow \Box p$ is false. So we start our tree by assuming that the *negation* of $\Diamond p \rightarrow \Box p$ is *true* at w. We write this down as follows.

1.	$\neg(\Diamond p \rightarrow \Box p)$	(w)	(A)

'1.' and '(A)' are for book-keeping; 'A' is short for 'Assumption', since we're *assuming* that $\neg(\Diamond p \rightarrow \Box p)$ is true at w. Now we unfold this assumption, by considering what the falsity of $\Diamond p \rightarrow \Box p$ at w implies for the two subsentences $\Diamond p$ and $\Box p$. By definition 2.2, a conditional $A \rightarrow B$ is false at w iff A is true at w and B is false. So $\Diamond p$ must be true at w while $\Box p$ is false. We expand the tree by adding these consequences.

1.	$\neg(\Diamond p \rightarrow \Box p)$	(w)	(A)	✓
2.	$\Diamond p$	(w)	(1)	
3.	$\neg\Box p$	(w)	(1)	

I have ticked off node 1 (with '✓') to mark that we won't need to look at it again, since all the information in node 1 is contained in nodes 2 and 3. The parenthetical '(1)' at nodes 2 and 3 reminds us that these assumptions are derived from node 1.

We continue drawing out further consequences. What does the truth of $\Diamond p$ at w imply for the subsentence p? By definition 2.2, there must be some world – let's call it v – at which p is true.

1.	$\neg(\Diamond p \rightarrow \Box p)$	(w)	(A)	✓
2.	$\Diamond p$	(w)	(1)	✓
3.	$\neg\Box p$	(w)	(1)	
4.	p	(v)	(2)	

Node 3 claims that $\Box p$ is false at w. By definition 2.2, $\Box p$ is true at w iff p is true at all worlds. So if $\Box p$ is false at w, there must be some world at which p is false. Let's introduce such a world, naming it u. Our tree looks as follows.

1.	$\neg(\Diamond p \rightarrow \Box p)$	(w)	(A)	✓
2.	$\Diamond p$	(w)	(1)	✓
3.	$\neg\Box p$	(w)	(1)	✓
4.	p	(v)	(2)	
5.	$\neg p$	(u)	(3)	

Now the only unprocessed nodes are assumptions about sentence letters and negations of sentence letters. Sentence letters don't have (non-trivial) subsentences, so there are no more assumptions to unpack. The tree is complete, and defines a countermodel for $\Diamond p \to \Box p$.

Let's read off the countermodel. There are three worlds in our tree: w, v, and u. So $W = \{w, u, v\}$. By node 4, p is true at v, so $V(p, v) = 1$. By node 5, p is false at u, so $V(p, u) = 0$. We don't know whether p is true or false at w, and it doesn't matter (otherwise the tree would say). As you can verify, $\Diamond p \to \Box p$ is indeed false at world w in any model in which $W = \{w, u, v\}$ and $V(p, u) = 0$ and $V(p, v) = 1$.

One more example, before I state the general rules. Let's try to find a countermodel for $\Box(p \to q) \to (p \to \Box q)$. That's another conditional, so we begin much like before.

1.	$\neg(\Box(p \to q) \to (p \to \Box q))$	(w)	(A)	✓
2.	$\Box(p \to q)$	(w)	(1)	
3.	$\neg(p \to \Box q)$	(w)	(1)	

Node 1 assumes that the negation of the conditional is true at some world w. Nodes 2 and 3 break down this assumption, using the fact that $\neg(A \to B)$ is true (at a world) iff A is true and B false. We could deal with node 2 next, but it's better to ignore it for the moment and process 3 first, which is yet another negated conditional.

4.	p	(w)	(3)
5.	$\neg\Box q$	(w)	(3)

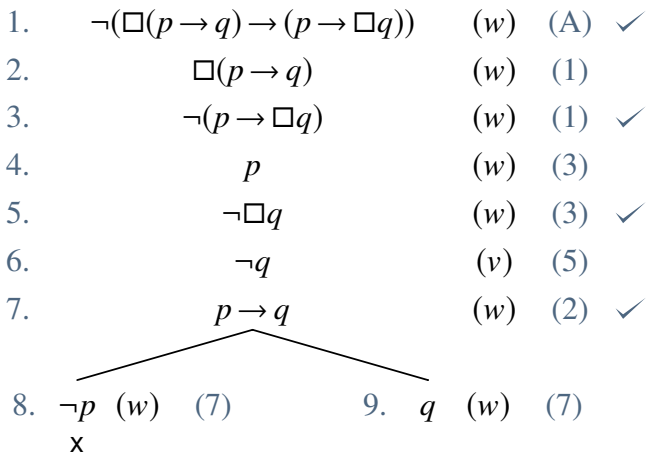
Node 5 tells us that q is not necessary (at w), so there is some world – call it v – at which q is false.

6.	$\neg q$	(v)	(5)

Now we need to return to node 2. What can we infer from the hypothesis that $\Box(p \to q)$ is true at w about the subsentence $p \to q$? By definition 2.2, $p \to q$ must be true at *every* world. So, in particular, $p \to q$ must be true at w. Let's write that down. We'll add another node for v later, so we don't check off node 2.

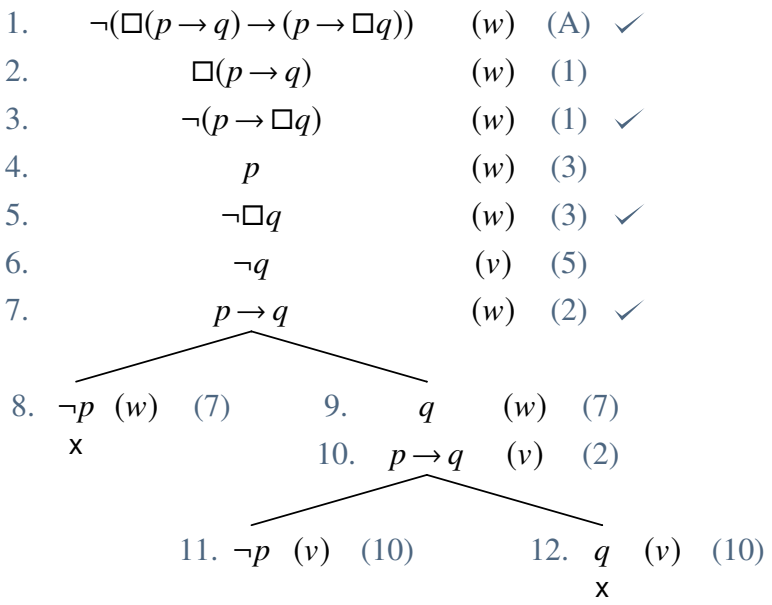
7.	$p \to q$	(w)	(2)

At this point, we face a problem. What can we infer from the truth of $p \to q$ at w about the subsentences p and q? By definition 2.2, $p \to q$ is true at w if *either* p is false at w *or* q is true at w. We have to keep track of both possibilities. So our (upside down) tree will branch. Here is the full tree at its present stage.



Recall that our goal is to construct a model in which the sentence at node 1 is true at world w. So far, the tree tells us that there are two worlds w and v in the model; lines 4 and 5 tell us something about the interpretation function in the model: p is true at w, q is false at v. After node 7, the tree branches, which means that it develops two ways of extending the model we have construed so far. On the left branch, we assume that p is false at w. On the right branch, we assume that q is true at w. But hold on: we already know that p is true at w (from node 4). There's no model in which p is both true and false at w. So the possibility explored on the left branch is a dead-end: it doesn't lead to a countermodel. That's why I've *closed* the branch by drawing a cross below node 8.

We continue on the right-hand branch. Here we expand node 2 again, this time for world v, which leads to another branching.



On the right-most branch, q is true at v (by node 12) but also false at v (by node 6), so that branch is closed. But the middle possibility is still open, and there are no more assumptions to unfold. So we have found a countermodel.

The countermodel has two worlds, $W = \{w, v\}$. The interpretation function V makes p true at w (node 4) and false at v (node 10); it makes q false at both v and w (nodes 6 and 9). In this model, $\Box(p \to q) \to (p \to \Box q)$ is false at w.

Now for the general rules.

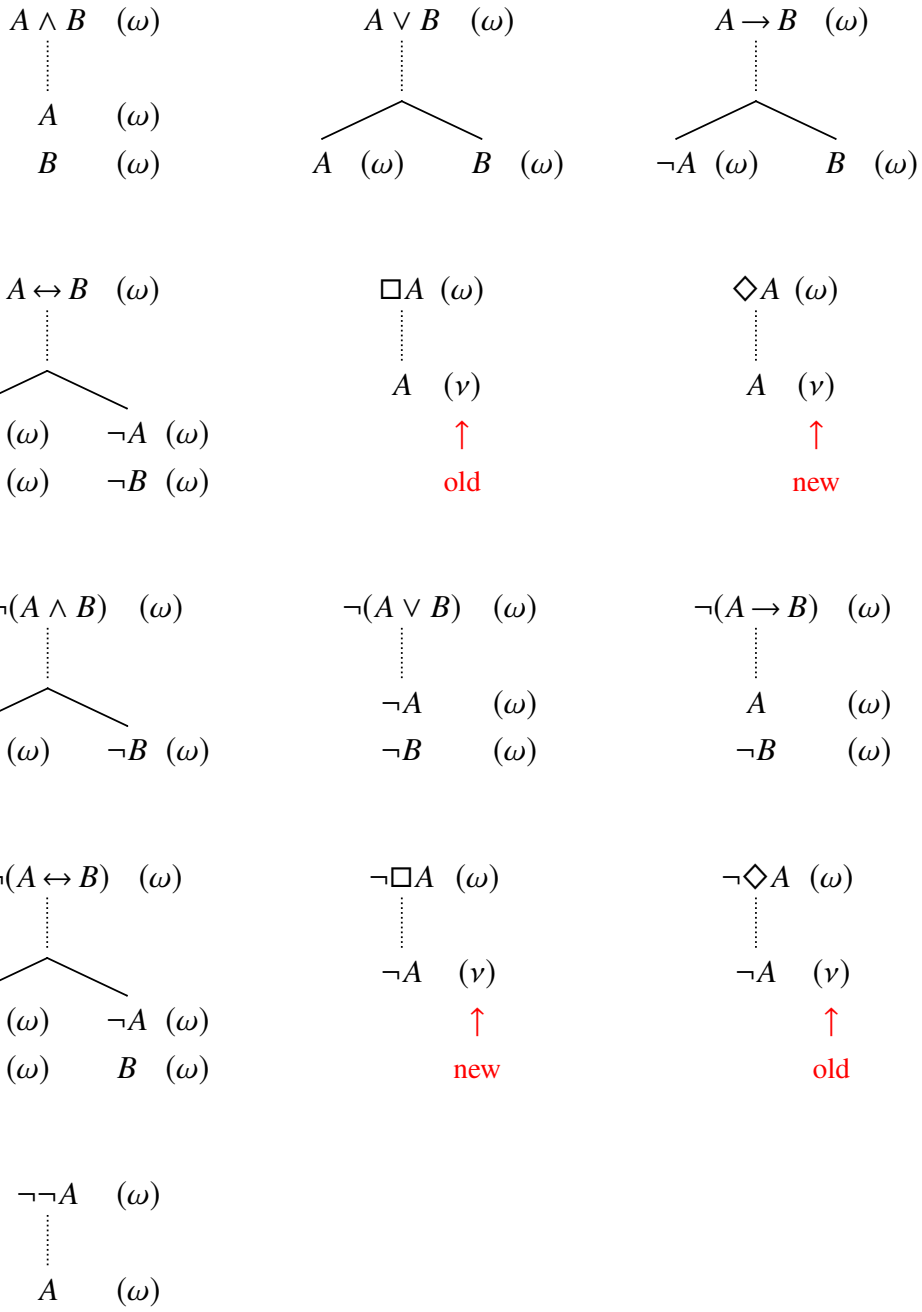
In order to find a countermodel for a sentence A with the help of the tree method, you always begin by assuming that the *negation* of A is true at world w:

1. $\neg A$ (w) (A)

You then expand this node, and every new node that appears on the tree, until no more nodes can be expanded.

To expand a non-negated node, you consider what the truth of the relevant sentence at the relevant world implies for the sentence's *immediate parts*. If the sentence has the form $A \wedge B$, $A \vee B$, $A \to B$, $A \leftrightarrow B$, then its immediate parts are the corresponding sentences A and B; sentences of the form $\Box A$, $\Diamond A$, and $\neg A$ have A as their only immediate part. To expand a negated node $\neg A$, you consider what the falsity of the negated sentence A at the relevant world implies for the immediate parts of A.

The following diagrams summarize how the different kinds of nodes are expanded.



If a branch of a tree (understood as extending all the way up to the root node) contains a sentence A as well as its negation $\neg A$, for the same world ω, then the branch is *closed* with an x at the bottom.

The rule for $\Box A$ says that from the assumption that $\Box A$ is true at a world ω, you may infer that A is true at any world *v that already occurs on the branch to which the new node is added*. So you're not allowed to introduce a new world variable when expanding $\Box A$ nodes. The same is true for $\neg\Diamond A$ nodes (which by duality means the same as $\Box\neg A$). By contrast, when you expand a $\Diamond A$ node (or a $\neg\Box A$ node), you must use a new world variable.

Nodes of type $\Box A$ and $\neg\Diamond A$ can be expanded several times, once for every world variable on any branch containing the node.

If you expand a node that is not of type $\Box A$ and $\neg\Diamond A$, the new nodes should be added to every open branch containing the node. The node can then be ticked off. $\Box A$ and $\neg\Diamond A$ nodes are never ticked off.

If no more rules can be applied, the tree is complete. Any open branch on the tree then defines a counterexample for the target sentence A.

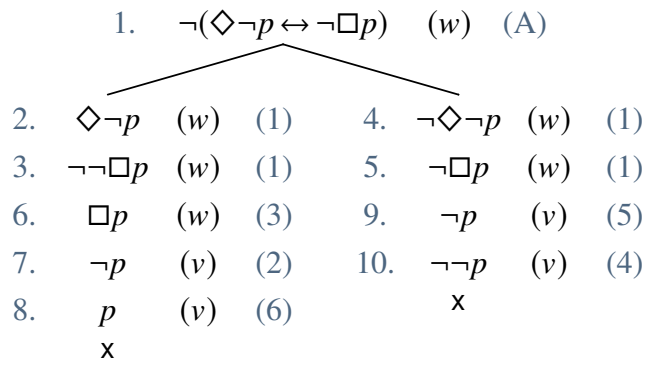
Exercise 2.7

Use the tree method to find countermodels for the following sentences:

(a) $p \to \Box(p \vee q)$
(b) $\Box p \vee \Box\neg p$
(c) $\Diamond(p \to q) \to (\Diamond p \to \Diamond q)$
(d) $p \to q$
(e) $\Box\Diamond p \to p$

What if all branches in a tree close? Then there is no countermodel for the target sentence. If there is no countermodel for a sentence, then the sentence is valid. This is how the tree method is used to show that sentences are valid.

For example, the following tree shows that $\Diamond\neg p \leftrightarrow \neg\Box p$ is valid. Make sure you understand each step. (I've omitted the check marks since these are only useful during the construction phase.)



A similar tree could obviously be drawn for $\Diamond\neg q \leftrightarrow \neg\Box q$, and for any other formula of the form $\Diamond\neg A \leftrightarrow \neg\Box A$: we would simply replace each occurrence of p on the tree with A. To show that a schema is valid, we can also directly draw **schematic trees** in which we use schematic variables 'A', 'B', 'C' instead of sentence letters.

Exercise 2.8

Use the tree method to show that the schemas **K**, **T**, **D**, **4**, **5**, and **G** are valid.

K $\Box(A \to B) \to (\Box A \to \Box B)$
T $\Box A \to A$
D $\Box A \to \Diamond A$
4 $\Box A \to \Box\Box A$
5 $\Diamond A \to \Box\Diamond A$
G $\Diamond\Box A \to \Box\Diamond A$

Exercise 2.9

For each of the following sentences, either show that it is valid or give a countermodel to show that it is invalid:

(a) $p \to \Box\Diamond p$
(b) $\Diamond\Diamond p \to \Diamond p$
(c) $\Diamond(p \wedge q) \to (\Diamond p \wedge \Diamond q)$
(d) $(\Diamond p \wedge \Diamond q) \to \Diamond(p \wedge q)$
(e) $\Diamond(p \vee q) \leftrightarrow (\Diamond p \vee \Diamond q)$
(f) $\Box\Diamond p \to \Diamond\Box p$

3 Accessibility

3.1 Variable modality

In the last chapter, we assumed that the box and the diamond quantify unrestrictedly over all possible worlds in a model. This has the consequence that modal sentences do not change their truth-value from world to world: if a sentence of the form $\Box A$ (or $\Diamond A$) is true at some world in a model, then it is automatically true at all worlds in the model.

For some interpretations of the box and the diamond, this is fine, but for many others, it is not. Suppose we read the box as 'it is known that'. Whether something is known varies from world to world. In some worlds, it is known who murdered Richard Montague, in others it is not. Or suppose we read the box as 'obligatory'. Again, what is obligatory plausibly varies from world to world. In worlds where you have promised to cook dinner, you are under an obligation to cook dinner; in other worlds, you do not have that obligation.

More obviously, suppose we read the box as 'from now on, it is always going to be the case that'. In models of temporal logic, the "worlds" W are interpreted as times. On the present interpretation, $\Box p$ is true at a given time t iff p is true at *all times after* t; times before t are irrelevant. So the box does not quantify unrestrictedly over all times. Moreover, the times over which it quantifies depend on the original time t: evaluated at time t, $\Box p$ quantifiers over times after t.

In this chapter, we will generalise the semantics from the previous chapter to allow for applications like these. The generalisation is easy. Intuitively, we simply assume that for any world w in W, there is a set of worlds that are possible *relative to* w; $\Box p$ is true at w iff p is true at all worlds that are possible relative to w. If a world v is possible relative to w we also say that v is *accessible from* w, or (informally) that w *can see* v.

On the new semantics, a model for $\mathfrak{L}_M$ has to specify which worlds in a model are accessibility from which other worlds, and from themselves. This marks the

difference between a "basic model" and a "Kripke model".

Definition 3.1

A **Kripke model** of $\mathfrak{L}_M$ is a triple $\langle W, R, V \rangle$ consisting of

- a non-empty set W,
- a binary relation R on W, and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value.

(R is called a relation *on* W because it relates only members of W.)

We also need to update definition 2.2, which settles under which conditions an $\mathfrak{L}_M$-sentence is true at a world in a model. The old definition had the following clauses for the box and the diamond:

(g) $M, w \models \Box A$ iff $M, v \models A$ for all $v \in W$.
(h) $M, w \models \Diamond A$ iff $M, v \models A$ for some $v \in W$.

In the new semantics, the box and the diamond only quantify over accessible worlds:

(g) $M, w \models \Box A$ iff $M, v \models A$ for all $v \in W$ such that wRv.
(h) $M, w \models \Diamond A$ iff $M, v \models A$ for some $v \in W$ such that wRv.

Here is the full definition, for completeness.

Definition 3.2: Kripke Semantics

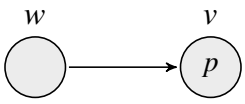
If $M = \langle W, R, V \rangle$ is a Kripke model, w is a member of W, ρ is any sentence letter, and A, B are any $\mathfrak{L}_M$-sentences, then

(a)	$M, w \models \rho$	iff $V(\rho, w) = 1$.
(b)	$M, w \models \neg A$	iff $M, w \not\models A$.
(c)	$M, w \models A \wedge B$	iff $M, w \models A$ and $M, w \models B$.
(d)	$M, w \models A \vee B$	iff $M, w \models A$ or $M, w \models B$.
(e)	$M, w \models A \rightarrow B$	iff $M, w \models B$ or $M, w \not\models A$.
(f)	$M, w \models A \leftrightarrow B$	iff $M, w \models (A \rightarrow B)$ and $M, w \models (B \rightarrow A)$.
(g)	$M, w \models \Box A$	iff $M, v \models A$ for all $v \in W$ such that wRv.
(h)	$M, w \models \Diamond A$	iff $M, v \models A$ for some $v \in W$ such that wRv.

When I speak of truth at a world in a Kripke model, this should always be understood in accordance with definition 3.2. Definition 2.2 defines truth at a world in a basic model.

Like definition 2.2, definition 3.2 settles the truth-value of any $\mathfrak{L}_M$-sentence at any world in any (relevant) model. Let's go through a few examples.

Consider a model with two worlds, w and v; v is accessible from w, and no world is accessible from v; the interpretation function assigns 1 to p at v and 0 to all other sentence letters and worlds. The model can be pictured like this, with an arrow representing accessibility:



With the help of definition 3.2, we can figure which $\mathfrak{L}_M$-sentences are true at which worlds in the model. For example:

- By clause (a) of definition 3.2, p is true at v and false at w.
- By clause (g), $\Box p$ is true at w, because p is true at v and v is the only world accessible from w.
- By clause (h), $\Diamond p$ is true at w, because p is true at v and v is accessible from w.
- By clause (h), $\Diamond p$ is false at v, because there is no world accessible from v at which p is true.
- By clause (g), $\Box\Diamond p$ is false at w, because $\Diamond p$ is false at v and v is accessible from w.

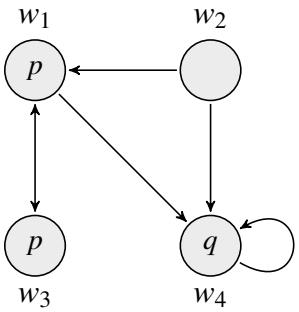
- By clause (g), $\Box\Diamond p$ is true at v, because there is no world accessible from v at which $\Diamond p$ is false.

Note that $\Diamond p$ and $\Box\Diamond p$ do not always have the same truth-value. In the new semantics, we can no longer ignore all but the last operator in a string of operators. Note also that $\Box p$ is true at w even though p is false; so $\Box p \rightarrow p$ is no longer valid.

Exercise 3.1

Explain why any sentence of the form $\Box A$ is true at world v in the above model.

Another example:



Using definition 3.2, we can figure out the following (among other things).

- $\Box p$ is false at w_1, because w_1 can see w_4, where p is false.
- $\Box p$ is true at w_3, because w_3 can only see w_1, where p is true.
- $\Diamond\Box p$ is true at w_1, because w_1 can see w_3 and $\Box p$ is true at w_3.
- $\Diamond q$ is true at w_4, because w_4 can see itself, and q is true at w_4.
- $\Diamond\Diamond q$ is true at w_1, because w_1 can see w_4, and $\Diamond q$ is true at w_4.

The next three exercises all refer to this example.

Exercise 3.2

At which worlds in the model are the following sentences true?

(a) $p \vee \neg q$
(b) $\Box(p \vee \neg q)$
(c) $\Diamond(\neg p \wedge \neg q)$

(d) $\Diamond\Box q$
(e) $\Diamond\Diamond\Box q$

Exercise 3.3

For each world in the model, find an $\mathfrak{L}_M$-sentence that is true only at that world.

Exercise 3.4

Can you draw a diagram of a smaller model (with fewer worlds) in which the exact same $\mathfrak{L}_M$-sentences are true at w_1?

3.2 The systems K and S5

Remember that a sentence is *valid* iff it is true at all worlds in all models. Different conceptions of a model, and of what it means for a sentence to be true at a world in a model, give rise to different kinds of validity. For example, $\Box p \rightarrow p$ is valid by the definitions from the previous chapter, but not by our present definitions.

To avoid confusion, it is best to use different expressions for the different kinds of validity. Let's call the new kind of validity *K-validity*. The old kind will henceforth be called *S5-validity*, because the sentences that are valid by the definition from the previous chapter are precisely the sentences contained in C.I. Lewis's system S5.

Definition 3.3

A sentence A is **K-valid** (for short, $\models_K A$) iff A is true at every world in every Kripke model.

The same distinction applies to the concept of entailment or logical consequence. Logical consequence in the old sense (definition 2.3) will henceforth be called *S5-consequence*. The new sense is that of *K-consequence*.

Definition 3.4

A sentence A is a **K-consequence** of a set Γ of sentences (for short, $\Gamma \models_K A$) iff A is true at any world in any Kripke model at which all members of Γ are true.

As before I'll also apply these notions to schematic sentences. For example, a schema is K-valid iff all its instances are K-valid.

Many properties of S5-consequence and S5-validity carry over to K-consequence and K-validity. In particular, observation 1.1 (p. 11), the propositional extension theorem (p. 33), and the replacement theorem (p. 33) still hold, and for the same reasons as before. I won't go through the arguments again.

Consider the set of all S5-valid sentences – that is, the set of sentences that are valid by the definitions of the previous chapter. This set is known as **system S5**. The set of all K-valid sentences is known as **system K**. Intuitively, these sets are called "systems" because they aren't random collection of sentences; the sentences in K or S5 are systematically related to one another. Systems are also often called **logics**.

You might think a logic should do more than identify a class of valid sentences: it should also tell us which sentences follow from which others. But our systems actually do that, albeit in an indirect manner. By observation 1.1, we can convert questions about logical consequence into questions about validity. For example, instead of asking whether $\Box p$ entails p, we can equivalently ask whether $\Box p \to p$ is valid. The systems S5 and K answer all questions of the latter kind, and so they indirectly settle whether $\Box p$ entails p. (Yes for S5, no for K.)

From the previous chapter, we know that S5 contains

- all instances of the **K**-schema $\Box(A \to B) \to (\Box A \to \Box B)$,
- all instances of the **T**-schema $\Box A \to A$,
- all instances of the **D**-schema $\Box A \to \Diamond A$,
- all instances of the **4**-schema $\Box A \to \Box\Box A$,
- all instances of the **5**-schema $\Box A \to \Box\Diamond A$, and
- all instances of the **G**-schema $\Diamond\Box A \to \Box\Diamond A$.

Which of these do we have in system K?

As we saw above, we do not have all instances of the **T**-schema, for there are worlds in Kripke models at which $\Box p \rightarrow p$ is false. So the **T** schema is not K-valid. Nor are the schemas **D**, **4**, **5**, and **G**.

Exercise 3.5

Can you find an instance of the **T** schema that is K-valid?

Exercise 3.6

Give a countermodel to **D**. That is, define a Kripke model in which $\Box p \rightarrow \Diamond p$ is false at some world w.

The **K** schema, however, *is* K-valid, as its name suggests.

Observation 3.1: Every instance of **K** is true at every world in every Kripke model.

Proof: Let w be an arbitrary world in an arbitrary Kripke model. By clause (e) of definition 3.2, an instance of $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ is false at w only if $\Box(A \rightarrow B)$ and $\Box A$ are both true at w while $\Box B$ is false. By clause (g) of definition 3.2, $\Box B$ is false at w only if B is false at some world accessible from w. But since $\Box(A \rightarrow B)$ and $\Box A$ are both true at w, $A \rightarrow B$ and A would have to be true at any such world, again by clause (g). But there can be no world at which $A \rightarrow B$ and A are true while B is false, by clause (e) of definition 3.2. □

Exercise 3.7

Show that $\Box(A \vee \neg A)$ is K-valid.

We have found sentences that are in S5 but not in K. Are there also sentences in K that are not in S5? You may want to pause and think about this for a moment.

Observation 3.2: Every K-valid sentence is S5-valid.

Proof: In essence, observation 3.2 holds because the basic models from the previous chapter can be simulated by Kripke models in which all worlds have universal access to all worlds. If a sentence is K-valid and hence true in every Kripke model, it must also true in every "universal" Kripke model; these models behave just like basic models; so the sentence is also true in every basic model.

It is worth going through this more carefully. For any basic model $M = \langle W, V \rangle$, let M^* be the Kripke model $\langle W, R, V \rangle$ with the same worlds W and the same interpretation function V, and with an accessibility relation R that holds between all worlds in W; so every world in M^* can see every other world as well as itself. If every world can see every world, then it makes no difference whether we use definition 2.2 or definition 3.2 to evaluate the truth of sentences at a world. That's because the two definitions only differ for the case of the modal operators, which definition 2.2 interprets as quantifiers over all worlds, while definition 3.2 interprets them as quantifiers over the accessible worlds. So we have:

(*) A sentence is true at a world w in a basic model M iff it is true at w in the corresponding Kripke model M^*.

(A full proof of (*) would proceed by induction on complexity of the sentence.) Now suppose a sentence A is *not* S5-valid, meaning that it is false at some world w in some basic model M. By (*), it follows that A is also false at some world in some Kripke model (namely, in M^*). And if A is false at some world in some Kripke model, then A is not K-valid. By contraposition, this means that if A is K-valid, then A is S5-valid. □

3.3 Some other normal systems

For many applications of modal logic, we need a concept of validity that lies in between K-validity and S5-validity. For example, suppose we read the box as physical necessity and the diamond as physical possibility, understood as compatibility with the laws of nature. On a certain conception of the laws of nature, the laws at a world cannot be violated at that world: anything that happens must be compatible with the laws. Equivalently, anything that is physically necessary is actually the case. In the logic of physical necessity, $\Box A$ should therefore entail A. On the other hand, it is not clear if $\Box A$ should entail $\Box\Box A$: if A is physically necessary, can we infer that it is physically necessary that A is physically necessary? Some have argued that we can't.

If that is right, then the logic of physical necessity lies in between K and S5. We want $\Box A \rightarrow A$ to be valid, but not $\Box A \rightarrow \Box\Box A$. System K gives us neither, S5 gives us both.

Fortunately, it is easy to define systems in between K and S5, by putting restrictions on the accessibility relation in Kripke models.

Let's say that an $\mathfrak{L}_M$-sentence is **valid in a class of Kripke models** iff the sentence is true at every world in every model in the class. (A schema is valid in a class of models iff all its instances are valid in the class.)

So K-validity is validity in the class of all Kripke models. $\Box p \rightarrow p$ is not K-valid, because it is false at certain worlds in certain Kripke models. All these worlds have something in common: they do not have access to themselves. If we rule out models in which worlds are inaccessible from themselves, the **T**-schema becomes valid.

Observation 3.3: The **T**-schema is valid in the class of Kripke models in which every world is accessible from itself.

Proof: According to clause (e) of definition 3.2, $\Box A \rightarrow A$ is false at a world w only if $\Box A$ is true at w and A is false; but if $\Box A$ is true at w and w has access to itself, then by clause (g) of definition 3.2, A is true at w. So if $\Box A \rightarrow A$ is false at w, and w is accessible from itself, then A is both true and false at w, which is impossible. Hence $\Box A \rightarrow A$ is true at every world in every model in which every world is accessible from itself. □

A relation R on a set W is called **reflexive** if each member of W is R-related to itself. If the accessibility relation in a Kripke model is reflexive, we'll also call the model itself reflexive. So observation 3.3 states that the **T**-schema is valid in the class of reflexive Kripke models.

The set of all sentences valid in the class of reflexive Kripke models is known as **system T**. System T is *stronger* than K, in the sense that it contains sentences not in K, but it is *weaker* than S5: it does not contain all of S5.

Systems of modal logic often share their name with a particular schema. To avoid confusion, I generally use bold-face letters for schemas and normal-face letters for systems. K and T are systems, **K** and **T** are schemas. All instances of **T** are in T, but many sentences in T (for example, all instances of **K**) are not instances of **T**.

Above I mentioned a temporal application of modal logic, in which the box is read

as 'it is always going to be the case that'. Here, $\Box p$ should count as true at a given time t iff p is true at all times *after* t. In the relevant Kripke models, the accessibility relation R is the earlier-later relation between times: $t_1 R t_2$ iff t_1 is earlier than t_2. On that interpretation, we don't want to assume that R is reflexive, which would mean that every point in time is earlier than itself. But we'll want something else. Suppose t_1 is earlier than t_2, and t_2 is earlier than t_3. Then surely t_1 is earlier than t_3. That is, we should restrict the relevant models to ones in which the accessibility relation is transitive.

A relation R is called **transitive** if, whenever xRy and yRz then xRz. Again, we will call a Kripke model transitive if its accessibility relation is transitive.

The set of sentences that are valid in the class of transitive Kripke models is known as **system K4**. The label alludes to the fact that (a) K4 is an extension of K, and (b) the **4**-schema is K4-valid. (a) is obvious. Here is an argument for (b):

Observation 3.4: The **4**-schema is valid in the class of transitive Kripke models.

Proof: Suppose for reductio that there is some transitive Kripke model in which some instance of $\Box A \rightarrow \Box\Box A$ is false at some world w. By clause (e) of definition 3.2, it follows that (1) $\Box A$ is true at w and (2) $\Box\Box A$ is false at w. By clause (g) of definition 3.2, (2) implies that there is some world v accessible from w where $\Box A$ is false. And that, in turn implies that there is some world u accessible from v at which A is false. Since R is transitive, u is accessible from w. By (1), A is true at u. So A is both true and false at u. Contradiction. □

We can combine the systems T and K4 by requiring both reflexivity and transitivity. The set of sentences valid in the class of reflexive and transitive Kripke models is C.I. Lewis's **system S4**. Both **T** and **4** are S4-valid.

There are many other conditions we could impose on the accessibility relation, and therefore many other systems of modal logic. Here are some well-known model classes with the conventional names for the corresponding systems, repeating (for future reference) the ones we already know.

System	*Constraint on R*
K	–
T	R is **reflexive**: every world in can access itself
D	R is **serial**: every world in can access some world
K4	R is **transitive**: whenever wRv and vRu, then wRu
K5	R is **euclidean**: whenever wRv and wRu, then vRu
B	R is reflexive and **symmetric**: whenever wRv then vRw
S4	R is reflexive and transitive
S4.2	R is reflexive, transitive, and **convergent**: whenever wRv and wRu, then there is some t such that vRt and uRt
S5	R is reflexive, transitive, and symmetric
S5	R is **universal**: every world has access to every world

We will have a closer look at some of these systems in later chapters, when we turn to applications of modal logic.

Any system that can be defined by putting constraints on the accessibility relation in Kripke models is called **normal**. So K, T, D, K4, K5, B, S4, and S5 are examples of normal systems, or normal logics. There are also non-normal systems/logics. These require a different kind of semantics. We will look at one alternative in section 6.5, but mostly we will stay within the realm of the normal.

S5 occurs twice in the above list. We already know S5 as the system for universal models, in which the box and the diamond quantify unrestrictedly over the whole space W. But we also get S5 if we require the accessibility relation to be reflexive, transitive, and symmetric.

Relations that are reflexive, transitive, and symmetric are called **equivalence relations**. An equivalence relation on a set W divides the members of W into classes within which the relation is everyone stands in the relation to everyone. (These classes are called **equivalence classes**.)

For example, let S be the relation that holds between two people if they have the same birthday. This is an equivalence relation: it is reflexive (everyone has the same birthday as themselves), transitive (if aSb and bSc than aSc), and symmetric (if aSb then bSa). for any person a, consider the class $[a]_S$ of everyone who has the same birthday as a. Everyone in $[a]_S$ has the same birthday as everyone else in $[a]_S$. So

within $[a]_S$, the same-birthday relation S is universal.

Now let me explain why the above two characterisations of S5 are equivalent.

Observation 3.5: A sentence is valid in the class of Kripke models whose accessibility relation is universal iff it is valid in the class of Kripke models whose accessibility relation is an equivalence relation.

Proof: The right-to-left direction is easy. If R is the universal relation on W, then R is reflexive, transitive, and symmetric. So the universal relations are a special kind of equivalence relation. If a sentence is valid in every model in which R is an equivalence relation, it must therefore be valid in every model in which R is universal.

The other direction is more interesting. We argue by contraposition, showing that if a sentence A is not valid in the class of models in which R is an equivalence relation, then R is also not valid in the class of universal models. So suppose A is not valid in the class of models in which R is an equivalence relation. Then there is some world w in some such model $M = \langle W, R, V \rangle$ such that $M, w \not\models A$. Define the new model $M' = \langle W', R', V' \rangle$ as follows:

> W' is the class of worlds accessible in M from w (i.e., the equivalence class $[w]_R$).
>
> R' is the universal relation on W'.
>
> V' assigns the same truth-value as V to all sentence letters within W'.

M' has a universal accessibility relation. But from the perspective of w, M and M' are indistinguishable. What (if anything) happens outside of $[w]_R$ in M makes no difference to the truth-value of any sentence at w: any sentence is true at w in M iff it is true at w in M'. This could be shown by induction on complexity, but I hope you see intuitively why it is the case.

From the assumption that A is false at some model whose accessibility relation is an equivalence relation, we can therefore infer that A is false in some model whose accessibility relation is universal. □

Exercise 3.8

Let R be the relation on the set of people that holds between x and y iff y is at least as old as x. Is R reflexive? serial? transitive? euclidean? symmetric? universal?

Exercise 3.9

Explain these facts:

(a) If R is symmetric and transitive, then R is euclidean.
(b) If R is symmetric and euclidean, then R is transitive.
(c) If R is reflexive and euclidean, then R is symmetric.

Exercise 3.10

What is wrong with the following argument? "If R is symmetric, then wRv implies vRw; if R is transitive, it follows that wRw. So symmetry and transitivity together imply reflexivity."

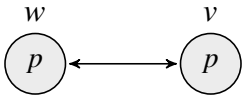
3.4 Frames

There is a close connection between conditions on the accessibility relation in Kripke models and modal schemas: between reflexivity and the **T** schema, between transitivity and the **4** schema, and so on. But what exactly is that connection?

You might think it's this:

(?) **T** is valid in a class of models iff all models in the class are reflexive; **4** is valid in a class of models iff all models in the class are transitive, and so on.

But that's false. Take the case of **T** and reflexivity. We know (observation 3.3) that **T** is valid in the class of reflexive models. It follows that if all models in a class are reflexive, then **T** is valid in that class. But the other direction fails. For a counterexample, consider the following model.



There are two worlds, both of which can see each other; neither can see itself. p is true at both worlds, all other sentence letters are false at both worlds. This model is not reflexive, but no instance of the **T** schema $\Box A \rightarrow A$ is false at any world in the model. (Try to find a false instance!) The fact that the **T** schema is valid in a class of models therefore does not entail that all models in the class are reflexive, for the class might contain models like the one just described.

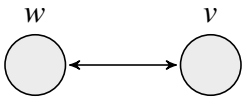
To understand the connection between modal schemas and conditions on the accessibility relation, we need to talk about *frames*. A frame is what you get if take a Kripke model and remove the interpretation function.

Definition 3.5

A **frame** is a pair of a set W and a relation R on W.

Roughly speaking, if we think of a model as a package of a scenario and an interpretation, then a frame is the logically relevant part of the scenario. What matters for logic is only the abstract structure of a scenario: how many worlds there are, and how they are accessible from one another.

Frames can be pictured just like Kripke models, but without any sentence letters in the nodes. The frame of the above model looks like this.



Now remember that validity is supposed to mean truth in virtue of the meaning of the logical expressions. Whether a sentence is valid should not depend on the meaning of the non-logical expressions. So if we define a particular kind of validity by reference to a class of Kripke models, the constraints we impose on the models in the class should be constraints on the frame of the models, not on the interpretation function.

For example, suppose I suggested that a sentence is "X-valid" iff it is true at all worlds in all Kripke model whose interpretation function assigns 1 to the sentence letter p at every world. Then $\Box p$ comes out X-valid, while $\Box q$ is X-invalid. Intuitively, X-validity is not a sensible concept of validity because $\Box p$ and $\Box q$ have the same logical form. If $\Box p$ is true in virtue of the meaning of the box, then $\Box q$ should also be true in virtue of the meaning of the box. The systems from the previous section

were all defined reasonably, by putting constraints on the frame of a Kripke model, rather than the interpretation function.

Let's say that a sentence is **valid on a frame** if it is true at all worlds in all models with that frame. A sentence is **valid in a class of frames** if it valid on all frames in the class.

Evidently, if a sentence is valid in the class of all models whose accessibility relation satisfies a certain condition, then it is also valid in the class of all frames whose accessibility relation satisfies that condition, and vice versa. So we could equivalently define system **T**, for example, as the set of sentences valid in the class of reflexive frames; **K4** is the set of sentences valid in the class of transitive frames; and so on. (A reflexive/transitive/etc. frame is a frame with a reflexive/transitive/etc. accessibility relation.)

Now here is the connection between **T** and reflexivity: **T** is valid in a class of frames iff all frames in the class are reflexive. More simply:

Observation 3.6: **T** is valid on a frame iff the frame is reflexive.

Proof: The right-to-left direction follows from observation 3.3, according to which **T** is valid in the class of reflexive models, and therefore in the class of reflexive frames, and therefore on any frame in that class. For the other direction, we have to show that if (all instances of) **T** are valid on a frame $\langle W, R \rangle$, then R is reflexive. We do this by showing that if R is not reflexive, then we can find an interpretation function V that makes $\Box p \rightarrow p$ false at some world w. w will be an arbitrary world in W that can't see itself. (There must be some such world if R is not reflexive.) We let V assign 0 to p at w and 1 to p at all other worlds. Then $\Box p$ is true at w and p false, and so $\Box p \rightarrow p$ is false at w. □

If a schema is valid on all and only the frames whose accessibility relation satisfies a certain property, the schema is said to **correspond** to that property. Observation 3.6 therefore says that the **T** schema corresponds to reflexivity.

Instead of proving more facts about the correspondence between modal schemas and frame conditions, I will simply give you a list of some important results.

Schema		*Corresponding Frame Condition*
T	$\Box A \rightarrow A$	R is reflexive: every world in W is accessible from itself
D	$\Box A \rightarrow \Diamond A$	R is serial: every world in W can access some world in W
B	$A \rightarrow \Box\Diamond A$	R is symmetric: whenever wRv then vRw
4	$\Box A \rightarrow \Box\Box A$	R is transitive: whenever wRv and vRu, then wRu
5	$\Diamond A \rightarrow \Box\Diamond A$	R is euclidean: whenever wRv and wRu, then vRu
G	$\Diamond\Box A \rightarrow \Box\Diamond A$	R is convergent: whenever wRv and wRu, then there is some t such that vRt and uRt

Correspondence facts are often useful when trying to figure out which schemas should be valid on a given interpretation of the modal operators. Return to the case of physical possibility and necessity. Above I asked whether the **4** schema $\Box A \rightarrow \Box\Box A$ should count as valid on this interpretation. The question is hard to answer by direct intuition: if something is physically necessary, is it physically necessary that it is physically necessary? Since the **4** schema corresponds to transitivity, we can equivalently ask whether the physical accessibility relation is transitive. That is, if a world v is physically possible relative to a world w, and u is physically possible relative to v, is u always physically possible relative to w?

The answer depends (among other things) on how we understand physical possibility. Earlier I suggested that a world v is physically possible relative to a world w if nothing that happens at v contradicts the laws of nature at w. This does not imply that v has the same laws as w. To illustrate, suppose the only law at w is that ravens are black; at v, there is no such law but there happen to be no non-black ravens. Then what happens at v does not contradict the laws at w, even though v has different laws. Relative to the laws of v, worlds with white ravens are physically possible. So a world accessible from a world that is accessible from w need not itself be accessible from w. So **4** is not valid in the logic of physical necessity.

Exercise 3.11

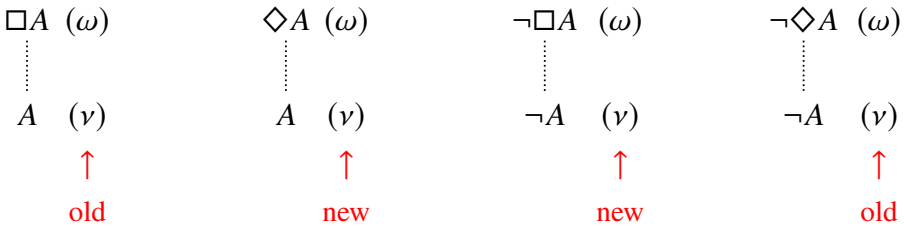
Can you find frame conditions that correspond to these schemas?

(a) $\Box A \leftrightarrow A$

(b) $\Box A$

3.5 More trees

In section 2.5, I described the tree method for checking whether a sentence or schema is valid, and for constructing countermodels. These were the rules for the box and the diamond:



The rule for $\Box A$ allows us to infer, from the hypothesis that $\Box A$ is true at some world, that A is true at any world that occurs on a tree branch. This made sense given the semantics of the previous chapter, where the box quantified unrestrictedly over all worlds. With the new semantics of the present chapter, we need to change the rules.

If $\Box A$ is true at a world w, and there's some other world v on the branch, we can only infer that A is true at v if v is accessible from w. So we need to keep track of which worlds are accessible from any world on a tree. We do this by adding meta-linguistic statements about accessibility to the tree.

For example, suppose we want to expand the following node.

$$\text{n.} \qquad \Diamond p \qquad (w)$$

The node represents the hypothesis that $\Diamond p$ is true at w. It follows that p is true at some world v. Moreover, that world v must be accessible from w. So we add two new nodes:

$$\begin{array}{lll} \text{m.} & wRv & \\ \text{m+1.} & p & (v) \end{array}$$

Node $m+1$ is what we would have added by the old rules. Node m is a meta-linguistic statement reminding us that v is accessible from w. 'wRv' is not a sentence of $\mathfrak{L}_M$; it isn't true or false relative to a world, which is why node m has no world label.

What if we want to expand a box node?

n. $\Box p$ (w)

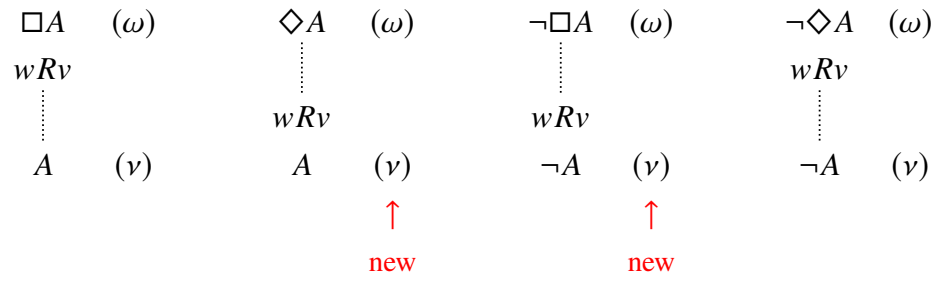
By itself, this doesn't tell us anything about the truth-value of p at any world. We can't infer that p is true at w, because w might not be accessible from itself. Indeed, if no world is accessible from w, then $\Box p$ can be true even if p is false at every world. So we can't even infer that there is some world or other at which p is true.

However, suppose a branch that contains node n also contains the following node.

m. wRv

Now we can infer that p is true at v. So to expand a box node on a branch, there must be another node on the branch telling us that the world at which the boxed sentence is true has access to some other world.

Here are diagrams of the new rules for the box and the diamond.



If two nodes occur above the dotted line in a rule, as in the rule for $\Box A$, this means that the rule can only be applied if both nodes already occur on the relevant branch (in any order, and not necessarily adjacent to each other).

The rules for negated boxes and diamonds are as you would expect from the duality of the box and the diamond. Note that only nodes of type $\Diamond A$ and $\neg\Box A$ allow us to introduce hypotheses about accessibility into a tree.

The rule for the classical connectives all stay the same. Together, all these rules are known as the **K-rules**; the tree rules from section 2.5 are the **S5-rules**.

Here is a schematic tree proof to show that $\models_K \Box(A \wedge B) \rightarrow (\Box A \wedge \Box B)$.

1.	$\neg(\Box(A \wedge B) \rightarrow (\Box A \wedge \Box B))$	(w)	(Ass.)				
2.	$\Box(A \wedge B)$	(w)	(1)				
3.	$\neg(\Box A \wedge \Box B)$	(w)	(1)				
4.	$\neg\Box A$	(w)	(3)	5.	$\neg\Box B$	(w)	(3)
6.	wRv		(4)	11.	wRu		(5)
7.	$\neg A$	(v)	(4)	12.	$\neg B$	(u)	(5)
8.	$A \wedge B$	(v)	(2,6)	13.	$A \wedge B$	(u)	(2,11)
9.	A	(v)	(8)	14.	A	(u)	(13)
10.	B	(v)	(8)	15.	B	(u)	(13)
	x				x		

The annotation '(2,6)' for node 8 indicates that this node is based on two assumptions from earlier in the branch: the assumption on node 2 that $\Box(A \wedge B)$ is true at w, and the assumption on node 6 that wRv. Only these two assumptions together allow us to infer that $A \wedge B$ is true at v.

Exercise 3.12

Use the K-rules to check which of the following schemas are K-valid.

(a) $(\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$
(b) $\Diamond(A \wedge B) \rightarrow (\Diamond A \wedge \Diamond B)$
(c) $(\Diamond A \wedge \Diamond B) \rightarrow \Diamond(A \wedge B)$
(d) $\Diamond(A \vee B) \leftrightarrow (\Diamond A \vee \Diamond B)$
(e) $\Box(A \vee B) \leftrightarrow (\Box A \vee \Box B)$
(f) $\Box(A \rightarrow B) \rightarrow (\Diamond A \rightarrow \Diamond B)$.
(g) $(\Box A \wedge \Diamond B) \rightarrow \Diamond(A \wedge B)$.

For normal systems in between K and S5 we exploit the fact that these systems can be defined in terms of constraints on the accessibility relation. So we simply add new rules for manipulating accessibility nodes on a tree, corresponding to the relevant constraints.

For example, if we want to check whether a sentence is T-valid, we add a *reflexivity*

rule to the K-rules. The reflexivity rule says that if a world variable ω occurs on a branch, then we may always add $\omega R\omega$ to the branch.

Here is a proof of $\Box p \to p$, using the reflexivity rule.

1.	$\neg(\Box p \to p)$	(w)	(Ass.)
2.	$\Box p$	(w)	(1)
3.	$\neg p$	(w)	(1)
4.	wRw		(Ref.)
5.	p	(w)	(2,4)
	x		

To test for validity on a class of transitive frames (or models), we need a *transitivity rule*, which allows us to infer $\omega R\upsilon$ from $\omega R\nu$ and $\nu R\upsilon$. Here is a proof of the **4** schema using this rule.

1.	$\neg(\Box A \to \Box\Box A)$	(w)	(Ass.)
2.	$\Box A$	(w)	(1)
3.	$\neg\Box\Box A$	(w)	(1)
4.	$wR\nu$		(3)
5.	$\neg\Box A$	(ν)	(3)
6.	νRu		(5)
7.	$\neg A$	(u)	(5)
8.	wRu		(4,6,Tr.)
5.	A	(u)	(2,8)
	x		

The following diagrams summarize the tree rules for the frame conditions we have so far considered.

Reflexivity	Seriality	Transitivity
⋮	⋮	$\omega R\nu$
$\omega R\omega$	$\omega R\nu$	$\nu R\upsilon$
↑	↑ ↑	⋮
old	old new	$\omega R\upsilon$

Symmetry	Euclidity	Convergence
$\omega R\nu$	$\omega R\nu$	$\omega R\nu$
⋮	$\omega R\upsilon$	$\omega R\upsilon$
$\nu R\omega$	⋮	⋮
	$\nu R\upsilon$	$\nu R\tau$
		$\upsilon R\tau$
		↑
		new

By selectively adding some of these rules to the K-rules, we get tree rules for a variety of normal modal logics, such as the following. (Compare the table on p. 54.)

System	*Tree Rules*
K	K-rules
T	K-rules and reflexivity rule
D	K-rules and seriality rule
K4	K-rules and transitivity rule
K5	K-rules and euclidity rule
B	K-rules and symmetry rule
S4	K-rules, reflexivity rule, and transitivity rule
S4.2	K-rules, reflexivity rule, transitivity rule, and convergence rule

4 Proofs about Proofs

4.1 Soundness and completeness

Recall the definition of K-validity from section 3.2. A sentence A is K-valid (for short, $\models_K A$) iff A is true at all worlds in all Kripke models. I have presented a tree method for figuring out whether a sentence is K-valid. But I haven't shown that this method *works*. By this I mean two things:

1. If all branches on a tree proof with the K-rules close, then the tested sentence is K-valid.

2. If the tested sentence is K-valid, then all branches on any tree proof with the K-rules eventually close.

In the present section, we will prove these two claims.

Claim 1 is known as **soundness**. In general, a proof method is *sound* with respect to a particular concept of validity if everything that is provable with the method is valid.

Claim 2 is known as **completeness**. A proof method is *complete* with respect to a concept of validity if everything that is valid is provable with the method.

Let's start with claim 1: soundness. I first outline the proof idea, then go through the details.

We want to show that if a tree for a target sentence A closes, and the tree conforms to the K-rules, then A is K-valid. We assume for reductio that the tree for A closes even though A is not K-valid. Our aim is to derive a contradiction.

If A is not K-valid, then $\neg A$ is true at some world w in some Kripke model M. Note that the closed tree begins with

$$1. \quad \neg A \quad (w)$$

So if we take 'w' to pick out w in M, then node 1 is a correct statement about M, in the sense that $\neg A$ is indeed true at w in M. Now we can show the following:

> *If all nodes on some branch of a tree are correct statements about M (relative to some interpretation of the world variables), and the branch is extended by the K-rules, then all nodes on at least one of the resulting branches are still correct statements about M.*

It follows that on some branch of the completed tree, all nodes are correct statements about M. But the tree is closed and so contains a pair of contradictory statements. These statements can't both be correct statements about M. There's our contradiction.

Let's go through this more slowly. First, let's say that a node on a tree is *a correct statement about* a model M relative to some function f that maps world variables to worlds in M iff either

- the node has the form $\omega R \nu$ and $f(\omega) R f(\nu)$, or
- the node has the form $A\ (\omega)$ and $M, f(\omega) \models A$.

We now prove the italicised statement above:

Soundness Lemma

If all nodes on some branch b of a tree are correct statements about M relative to some mapping from world variables to worlds in M, and the branch is extended by applying a K-rule, then all nodes on at least one of the resulting branches are still correct statements about M relative to some mapping from world variables to worlds in M.

Proof: We have to go through all the K-rules. In each case we assume that the rule is applied to some node(s) on some branch b, all nodes on which are correct statements about M relative to some mapping f. We show that on at least one of the resulting branches, the all nodes are still correct statements about M relative to some mapping.

- Suppose b contains a node of the form $A \wedge B\ (\omega)$ and the branch is extended by two new nodes $A\ (\omega)$ and $B\ (\omega)$. By assumption, $A \wedge B\ (\omega)$ is a correct statement about M relative to f; that is, $M, f(\omega) \models A \wedge B$. It follows that $M, f(\omega) \models A$

and $M, f(\omega) \models B$. So the newly added nodes are also correct statements about M relative to f.

- Suppose b contains a node of the form $A \vee B$ (ω) and the branch is split into two, with A (ω) appended to one end and B (ω) to the other. By assumption, $M, f(\omega) \models A \vee B$. It follows that either $M, f(\omega) \models A$ or $M, f(\omega) \models B$. So at least one of the newly added nodes is also a correct statement about M relative to f.
- The proof for the other non-modal rules is similar. Let's move on to the rules for modal operators.
- Suppose b contains nodes of the form $\Box A$ (ω) and $\omega R \nu$, and the branch is extended by adding A (ν). By assumption, $M, f(\omega) \models \Box A$. and $f(\omega) R f(\nu)$. It follows that $M, f(\nu) \models A$. So the newly added node is also a correct statement about M relative to f.
- Suppose b contains a node of the form $\Diamond A$ (ω) and the branch is extended by adding nodes $\omega R \nu$ and A (ν), where ν is new on the branch. By assumption, $M, f(\omega) \models \Diamond A$. It follows that $M, v \models A$ for some v in M such that $f(\omega) R \nu$. Let f' be the same as f except that $f'(\nu) = v$. Then the newly added nodes are correct statements about M relative to f'. Moreover, since f and f' differ at most about ν and ν didn't occur on b before the addition of $\omega R \nu$ and A (ν), all earlier nodes on b are also correct statements about M relative to f'.
- The cases for $\neg\Box$ and $\neg\Diamond$ are similar. □

Let's spell out the full proof.

Theorem: Soundness of the tree rules for K

If a tree for a target sentence A closes, and the tree conforms to the K-rules, then A is K-valid.

We assume for reductio that the tree for A closes even though A is not K-valid. If A is not K-valid, then $\neg A$ is true at some world w in some Kripke model M. The first node on the tree, $\neg A$ (w), is a correct statement about M. Since the completed tree

is created from this starting point by applying the K-rules, the Soundness Lemma implies that some branch on the tree only contains correct statements about M relative to some mapping from world variables to the worlds in M. But the tree closes, meaning that all its branches contain contradictory nodes of the form

n. B (υ)

m. $\neg B$ (υ)

These two nodes can't both be correct statements about M. So we've reached a contradiction from the assumption that the tree for A closes even though A is not K-valid. □

Exercise 4.1

Spell out the clauses for $A \rightarrow B$ and $\neg\Diamond$ in the proof of the soundness lemma.

Next, let's prove completeness. We want to show that whenever a sentence A is K-valid, then there is a closed tree for A.

The proof is by comtraposition. We suppose there is no closed tree for A. We show that there is some world w in some Kripke model M at which A is false. The model will be read off from any complete but open tree for A.

A tree is *complete* if every rule than can be applied has been applied. (A complete tree may be infinite.) For any branch b in any complete tree for A, we say that the model *induced* by this branch is defined so that

- W is the set of world variables on the branch,
- $\omega R\nu$ iff $\omega R\nu$ occurs on the branch,
- for any sentence letter ρ and world ω, $V(\rho, \omega) = 1$ if $\rho(\omega)$ occurs on the branch, otherwise $V(\rho, \omega) = 0$.

This is just how we read off countermodels from a branch. Now we show that all nodes on b are correct statements about the model induced from the branch (relative to the function that maps each world variable to itself).

Completeness Lemma

If b is an open branch of a complete tree, and $M = \langle W, R, V \rangle$ is the model induced by b, then for all sentences A,

> if A (ω) is on b then $M, \omega \models A$;
> if $\neg A$ (ω) is on b then $M, \omega \not\models A$.

The proof is by induction on complexity of A.

- If A is a sentence letter, then the claim is true by definition.
- If A is a conjunction $B \wedge C$, then the conjunction rule has been applied, resulting in two nodes B (ω) and C (ω) on b. By induction hypothesis, $M, \omega \models B$ and $M, \omega \models C$. So $M, \omega \models B \wedge C$.
- If A is a disjunction $B \vee C$, then the disjunction rule has been applied, and at least one of the resulting nodes B (ω) and C (ω) lies on b. By induction hypothesis, $M, \omega \models B$ or $M, \omega \models C$. So $M, \omega \models B \vee C$.
- The proof for the other non-modal rules is similar. Let's move on to the rules for modal operators.
- If A has the form $\Box B$, then B (ν) in on b for all ν for which $\omega R \nu$ is on b (because that's what the box rule allows, and we assume the branch is complete). By induction hypothesis, $M, \nu \models B$ for all such ν. By construction of M, it follows that $M, \nu \models B$ for all worlds ν such that $\omega R \nu$. So $M, \omega \models \Box B$.
- If A has the form $\Diamond B$, then $\omega R \nu$ and B (ν) are on b, for some ν. By induction hypothesis, $M, \nu \models B$, and by construction of M, $\omega R \nu$. So $M, \omega \models \Diamond B$. □

The remainder of the completeness proof is easy.

Completeness of the tree rules for K

If A is K-valid, then there is a closed tree for A conforming to the K-rules.

Proof: Suppose for reductio that there is no closed tree for A conforming to the K-rules. Take any branch on any complete open tree for A. By the completeness lemma, the model induced by that branch makes A false at w. So A is not true at all worlds in all models. □

Exercise 4.2

Spell out the clause for $A \rightarrow B$ in the proof of the completeness lemma.

So the tree rules for K *work*: they allow us to prove all and only the K-valid sentences.

The tree rules for S5 also work, as do the rules for the various other systems we have looked at. In each case, the soundness and completeness proofs are analogous to the proofs for K.

Exercise 4.3

When constructing a tree proof, one often has a choice of which rules to apply in which order. In principle, it doesn't matter: if any tree for the target sentence closes, then no complete tree for the target sentence remains open. Explain why!

4.2 Axiomatic proofs

The tree method is easy to use. The older method of axiomatic proofs is not. The method is nonetheless worth studying, and not only because of its historical importance.

An axiomatic ("Hilbert-style") proof is a sequence of sentences each of which is either an axiom or follows from earlier sentences in the sequence by a rule.

In order to reduce the required axioms and rules, it is customary to use a limited version of $\mathfrak{L}_M$ in which the only logical operators are $\neg$, $\rightarrow$, and $\Box$. This is no real loss, since any sentence with other operators can be transformed into an equivalent

sentence with only $\neg$, $\rightarrow$, and $\Box$, by the following equivalences:

$$\begin{aligned} A \wedge B &\Leftrightarrow \neg(A \rightarrow \neg B) \\ A \vee B &\Leftrightarrow \neg A \rightarrow B \\ A \leftrightarrow B &\Leftrightarrow \neg((A \rightarrow B) \rightarrow \neg(B \rightarrow A)) \\ \Diamond A &\Leftrightarrow \neg\Box\neg A \end{aligned}$$

A well-known axiomatic calculus for classical propositional logic consists of the following three axiom schemas, **A1**–**A3**, together with the rule of *Modus Ponens*, **MP**.

$$A \rightarrow (B \rightarrow A) \qquad (\mathbf{A1})$$

$$(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)) \qquad (\mathbf{A2})$$

$$(\neg A \rightarrow \neg B) \rightarrow (B \rightarrow A) \qquad (\mathbf{A3})$$

If A and $A \rightarrow B$ occur on a proof, you may append B. (**MP**)

Each axiom schema has infinitely many instances, all of which count as axioms.

Any sentence that is valid in classical propositional logic can be derived from some instances of A1–A3 with the use of MP. To illustrate, here is a proof of $p \rightarrow p$, with added annotations.

1.	$p \rightarrow ((p \rightarrow p) \rightarrow p)$	(A1)
2.	$(p \rightarrow ((p \rightarrow p) \rightarrow p)) \rightarrow ((p \rightarrow (p \rightarrow p)) \rightarrow (p \rightarrow p))$	(A2)
3.	$(p \rightarrow (p \rightarrow p)) \rightarrow (p \rightarrow p)$	(1, 2, MP)
4.	$p \rightarrow (p \rightarrow p)$	(A1)
5.	$p \rightarrow p$	(3, 4, MP)

To get a complete axiomatization for system K, we only need one more axiom schema and one more rule. The axiom schema is **K** (as you may have guessed).

$$\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B) \qquad (\mathbf{K})$$

The new rule is called "necessitation":

If A occurs on a proof, you may append $\Box A$. (**Nec**)

In general, of course, one can't infer $\Box A$ from A. After all, the sentence A may be true at a world while $\Box A$ is false. However, the **Nec** rule is not a rule for drawing inferences from arbitrary assumptions. It allows you to infer $\Box A$ only if A already occurs earlier in the same proof. And everything that occurs in an axiomatic proof is either an axiom or it follows from the axioms by the rules. Given that the axioms are logical truths, and given that whatever logically follows from a logical truth itself a logical truth, any sentence that occurs on some line in an axiomatic proof is a logical truth. In effect, **Nec** therefore assumes that if a sentence A is logically true, then $\Box A$ is also logically true.

To illustrate the use of **K** and **Nec**, let's look at a proof of $\Box(p \wedge q) \rightarrow \Box p$. In our restricted version of $\mathfrak{L}_M$, this is $\Box\neg(p \rightarrow \neg q) \rightarrow \Box p$. The proof begins with a proof of the tautology $\neg(p \rightarrow \neg q) \rightarrow p$, using **A1**–**A3** and **MP**. This takes about 60 lines, which we don't need to worry about. From then, we proceed as follows.

60.	$\neg(p \rightarrow \neg q) \rightarrow p$	(derived from A1–A3 by MP)
61.	$\Box(\neg(p \rightarrow \neg q) \rightarrow p)$	(60, Nec)
62.	$\Box(\neg(p \rightarrow \neg q) \rightarrow p) \rightarrow (\Box\neg(p \rightarrow \neg q) \rightarrow \Box p)$	(K)
63.	$\Box\neg(p \rightarrow \neg q) \rightarrow \Box p$	(61, 62, MP)

I will use the name 'AK' for the axiomatic calculus consisting of the axioms **A1**–**A3**, **K**, and the rules **MP** and **Nec**.

Other axiomatic calculi for modal logic can be defined by adding further axioms or rules to the axioms and rules of AK. All calculi that can be defined in this way are called *normal*. The following table gives a few examples.

Axiomatic Calculus	Axioms	Rules
AK	**A1**–**A3**, **K**	**MP**, **Nec**
AD	**A1**–**A3**, **K**, **D**	**MP**, **Nec**
AT	**A1**–**A3**, **K**, **T**	**MP**, **Nec**
AK4	**A1**–**A3**, **K**, **4**	**MP**, **Nec**
AK5	**A1**–**A3**, **K**, **5**	**MP**, **Nec**
AS4	**A1**–**A3**, **K**, **T**, **4**	**MP**, **Nec**
AS4.2	**A1**–**A3**, **K**, **T**, **4**, **G**	**MP**, **Nec**
AS5	**A1**–**A3**, **K**, **T**, **5**	**MP**, **Nec**

Exercise 4.4

Outline a proof of $\Box p \to \Diamond p$ in AT. You can assume that any propositional tautology is somehow derivable from **A1**–**A3** and **MP**, without giving its proof. But do fill in all other steps.

Exercise 4.5

Suppose we add the ("McKinsey") schema $\Box\Diamond A \to \Diamond\Box A$ to AS5. Explain why we can then prove all instances of the ("Triv") schema $\Box A \leftrightarrow A$. You can assume that any propositional tautology is provable in the calculus, and that if a sentence is a truth-functional consequence of some provable sentences, then the sentence is also provable.

Soundness of axiomatic calculi is usually easy to prove.

Theorem: Soundness of AK

If $\vdash_{\mathrm{AK}} A$ then $\models_K A$.

Proof: We have to show that anything that is derivable from (some instances of) **A1**–**A3** and **K** by **MP** and **Nec** is true at all worlds in all Kripke models. To this end, we first show that all the axioms are true at all worlds in all Kripke models.

Then we show that if a sentence A follows from other sentences by **MP** or **Nec**, and the other sentences are true at all worlds in all Kripke models, then A is also true at all worlds in all Kripke models. That is, we will show that the axioms are valid, and that the rules preserve validity. It follows that anything that can be derived from the axioms by the rules is valid.

So let's go through the axiom schemas.

A1 By clause (e) of definition 3.2, an instance of $A \to (B \to A)$ is *false* at some world w in some Kripke model M iff (the corresponding instances of) A and B are both true at w while A is false at w. But we can't have both $M, w \models A$ and $M, w \not\models A$. So the assumption that some instance of $A \to (B \to A)$ is false at some world in some Kripke model has led to a contradiction. So every instance of the schema is true at every world in every Kripke model.

A2 By clause (e) of definition 3.2, an instance of $(A \to (B \to C)) \to ((A \to B) \to (A \to C))$ is false at some world w in some Kripke model M iff (the corresponding instances of) $A \to (B \to C)$, $A \to B$, and A are true at w while C is false at w. However, clause (e) also implies that if $A \to B$ and A are both true at w, then so is B. Similarly, clause (e) implies that if $A \to (B \to C)$ and A and B are true at w, then so is C. We have reached a contradiction: $M, w \models C$ and $M, w \not\models C$. So every instance of $(A \to (B \to C)) \to ((A \to B) \to (A \to C))$ is true at every world in every Kripke model.

A3 Exercise.

K See observation 3.1 on p.51.

It remains to show that **MP** and **Nec** preserve validity.

MP By clause (e) of definition 3.2, if some instance of $A \to B$ is true at some world in some Kripke model, and so is (the corresponding instance of) A, then B is also true at that world in that model. So if $A \to B$ and A are true at *all* worlds in *all* Kripke models, then B is also true at all worlds in all Kripke models.

Nec By clause (g) of definition 3.2, $\Box A$ is true at a world w in a Kripke model iff A is true at all worlds accessible from w. It follows that if A is true at *all* worlds in all Kripke models, then $\Box A$ is also true at all worlds in all Kripke models.

□

Exercise 4.6

Fill in the missing case for **A3**.

We've shown that whenever $\vdash_{\mathrm{AK}} A$ then $\models_K A$. Analogous proofs can be given for other axiomatic calculi. I'll outline one more proof to illustrate the pattern.

Theorem: Soundness of AT

If $\vdash_{\mathrm{AT}} A$ then $\models_T A$.

Proof sketch: We have to show that anything that is derivable from (some instances of) **A1**–**A3**, **K**, and **T** by **MP** and **Nec** is true at all worlds in all reflexive Kripke models. We've just shown that all instances of **A1**–**A3** and **K** are true at all worlds in all Kripke models, so they are also true at all worlds in all reflexive Kripke models. Moreover, by observation 3.3, every instance of the **T**-schema is true at all worlds in all reflexive Kripke model. Finally, the above arguments for **MP** and **Nec** really show that these rules preserve validity in any class of Kripke models. So they preserve validity in the class of reflexive Kripke models. □

Exercise 4.7

Sketch the proof that AS4 is sound with respect to $\models_{S4}$.

4.3 Canonical models

Next, completeness. I am going to show that any K-valid sentence is derivable from some instances of **A1**–**A3** and **K** by the rules **MP** and **Nec**. The argument will be long and complicated, so make sure you are sitting comfortably (and don't worry if you don't immediately get all the details).

I will argue by contraposition. I will show that any sentence that *cannot* be derived from **A1**–**A3** and **K** by **MP** and **Nec** is *not* K-valid. To show that a sentence is not K-valid, I will give a countermodel: a Kripke model in which the sentence is false at some world. You might think we need different countermodels for different invalid sentences, but it turns out that there's a special Kripke model that provides a

model for *every* sentence that isn't provable in AK. This special model is called the *canonical model* for AK.

To proof completeness for AK, I will use some intermediate results ("lemmas"). These results hold not just for AK, but for every normal axiomatic calculus. In what follows, I will therefore use 'AX' as a variable for an arbitrary axiomatic calculus that contains the axioms **A1**–**A3**, **K**, and the rules **MP** and **Nec**, with possibly further axioms or rules.

Now I need some more terminology. A set of $\mathfrak{L}_M$-sentences is called *AX-inconsistent* if it contains a finite number of sentences $A_1, \ldots, A_n$ such that $\vdash_{\mathrm{AX}} \neg(A_1 \wedge \ldots \wedge A_n)$. A set is *AX-consistent* if it is not AX-inconsistent.

For example, the set $\{\Box(p \wedge q), q \rightarrow p, \neg\Box q\}$ is AK-inconsistent, because it contains two sentences, $\Box(p \wedge q)$ and $\neg\Box q$ whose conjunction is refutable in AX, in the sense that the negation $\neg(\Box(p \wedge q) \wedge \neg\Box q)$ of their conjunction is derivable from some instances of **A1**–**A3** and **K** by **MP** and **Nec**.

A set of sentences is *maximal* if it contains either A or $\neg A$, for every sentence A. A set is *maximal AK-consistent* if it is both maximal and AK-consistent.

Now we can define the canonical model for AK and other axiomatic calculi.

Definition 4.1: Canonical model

The **canonical model** M_{AX} for a normal axiomatic calculus AX is the Kripke model $\langle W, R, V \rangle$, where

- W is the set of all maximal AX-consistent sets of $\mathfrak{L}_M$-sentences.
- For any $w, v \in W$, wRv iff v contains every sentence A for which w contains $\Box A$.
- For every sentence letter ρ and world w, $V(\rho, w) = 1$ iff $\rho \in w$.

So the "worlds" in a canonical model are sets of $\mathfrak{L}_M$-sentences. That's OK: nothing in the definition of a Kripke model says that the members of W are not sets of sentences. The interpretation function makes a sentence letter true at a "world" iff the letter is an element of the world. As we are going to see, this generalizes to arbitrary sentences:

(1) A world w in M_{AX} contains all and only the sentences that are true at w in M_{AX}.

We will also prove the following:

(2) If some sentence cannot be proved in the axiomatic calculus AX, then its negation is a member of some world in M_{AX}.

Why is that useful? Recall that to prove the completeness of AK, we want to show that if a sentence A can't be proved in AK, then there is some world in some Kripke model at which A is false. Fact (2) tells us that if A can't be proved, then $\neg A$ is a member of some world w in the canonical model M_{AK}. By fact (1), we can infer that $\neg A$ is true at w in M_{AK}. So A is false at w in M_{AK}. Which proves that A is not K-valid. So once we have established (1) and (2), we have all but established the completeness of AK.

I am going to prove (2) first. I'll need the following observation.

Observation 4.1: For any normal axiomatic calculus AX, if a set Γ is AX-consistent, then for any sentence A, either $\Gamma \cup \{A\}$ or $\Gamma \cup \{\neg A\}$ is AX-consistent. (Recall that $\Gamma \cup \{A\}$ is the set that contains all members of Γ as well as A.)

Proof sketch: Suppose for reductio that both $\Gamma \cup \{A\}$ and $\Gamma \cup \{\neg A\}$ are AX-inconsistent.

If $\Gamma \cup \{A\}$ is AX-inconsistent, then (by definition of AX-inconsistency) there are sentences $A_1, \ldots, A_n$ in $\Gamma \cup \{A\}$ such that $\vdash_{\mathrm{AX}} \neg(A_1 \wedge \ldots \wedge A_n)$. Since Γ itself is AX-consistent, one of the sentences $A_1, \ldots, A_n$ must be A. Let B be the conjunction of the other sentences in $A_1, \ldots, A_n$, all of which are in Γ. So $\vdash_{\mathrm{AX}} \neg(B \wedge A)$.

Similarly, if $\Gamma \cup \{\neg A\}$ is not AX-consistent, there are sentences $A_1, \ldots, A_n$ in $\Gamma \cup \{\neg A\}$ such that $\vdash_{\mathrm{AX}} \neg(A_1 \wedge \ldots \wedge A_n)$, and one of these sentences must be $\neg A$. Let C be the conjunction of the others, all of which are in Γ. So $\vdash_{\mathrm{AX}} \neg(C \wedge \neg A)$.

Now in propositional logic, $\neg(B \wedge A)$ and $\neg(C \wedge \neg A)$ entail $\neg(B \wedge C)$. And indeed, $\neg(B \wedge C)$ is derivable from $\neg(B \wedge A)$ and $\neg(C \wedge \neg A)$ through some applications of **A1–A3** and **MP**. So $\vdash_{\mathrm{AX}} \neg(B \wedge C)$. But $B \wedge C$ is a conjunction of sentences from Γ. So Γ is AX-inconsistent. But we know that Γ is AX-consistent. So our hypothesis that both $\Gamma \cup \{A\}$ and $\Gamma \cup \{\neg A\}$ are AX-inconsistent is false. □

Now we can prove fact (2):

Lindenbaum's Lemma

For any normal axiomatic calculus AX, any AX-consistent set is included in some maximal AX-consistent set.

Proof: Let S_0 be some AX-consistent set of sentences. Let $A_1, A_2, \ldots$ be a list of all $\mathfrak{Q}_M$-sentences. For every number $i \geq 0$, define

$$S_{i+1} = \begin{cases} S_i \cup \{A_i\} & \text{if } S_i \cup \{A_i\} \text{ is AX-consistent} \\ S_i \cup \{\neg A_i\} & \text{otherwise.} \end{cases}$$

This gives us an infinite list of sets $S_0, S_1, S_2, \ldots$. We first note that each set in the list is AX-consistent. S_0 is AX-consistent by assumption. And if some set S_i in the list is AX-consistent, then either $S_i \cup \{A_i\}$ is AX-consistent, in which case $S_{i+1} = S_i \cup \{A_i\}$ is AX-consistent, or $S_i \cup \{A_i\}$ is not AX-consistent, in which case S_{i+1} is $S_i \cup \{\neg A_i\}$, which is AX-consistent by observation 4.1. So if any set in the list is consistent, then the next set in the list is also consistent. So $S_0, S_1, S_2, \ldots$ are all AX-consistent.

Now let S be the set of sentences that occur in at least one of the sets $S_1, S_2, S_3 \ldots$. (In set theory symbols, $S = \bigcup_i S_i$.)

Evidently, S_0 is included in S. And S is maximal. And S is AX-consistent. For if S were not AX-consistent, then it would contain some sentences $B_1, \ldots, B_n$ such that $\vdash_{\mathrm{AX}} \neg(B_1 \wedge \ldots \wedge B_n)$. All of these sentences occur somewhere on the list $A_1, A_2, \ldots$. Let A_j be a sentence from $A_1, A_2, \ldots$ that occurs after all the $B_1, \ldots, B_n$. If $B_1, \ldots, B_n$ are in S, they would have to be in S_j already, so S_j would be AX-inconsistent. But we've seen that all of $S_0, S_1, S_2, \ldots$ are AX-consistent. So S is a maximal AX-consistent set that includes S_0. □

To prove fact (1), I need another observation.

Observation 4.2: If Γ is a maximal AX-consistent set of sentences that does not contain $\Box A$, and Γ^- is the set of all sentences B for which $\Box B \in \Gamma$, then $\Gamma^- \cup \{\neg A\}$ is AX-consistent.

Proof: We show that if $\Gamma^- \cup \{\neg A\}$ is not AX-consistent, then neither is Γ. If $\Gamma^- \cup \{\neg A\}$ is not AX-consistent, then there are sentences $B_1, \ldots, B_n \in \Gamma^-$ such that $\vdash_{\mathrm{AX}} \neg(B_1 \wedge \ldots \wedge B_n \wedge \neg A)$. Since AX extends propositional logic, $\vdash_{\mathrm{AX}} (B_1 \wedge \ldots \wedge B_n) \rightarrow A$. By repeated application of **Nec**, **K**, and **MP**, then $\vdash_{\mathrm{AX}} (\Box B_1 \wedge \ldots \wedge \Box B_n) \rightarrow \Box A$. And then $\vdash_{\mathrm{AX}} \neg(\Box B_1 \wedge \ldots \wedge \Box B_n \wedge \neg \Box A)$, by propositional logic. So $\{\Box B_1, \ldots, \Box B_n, \neg \Box A\}$ is not AX-consistent. But $\{\Box B_1, \ldots, \Box B_n, \neg \Box A\}$ is a subset of Γ, because $\Box B_1, \ldots, \Box B_n$ are in Γ^- and $\neg \Box A$ is in Γ because $\Box A \notin \Gamma$ and Γ is maximal consistent. □

Canonical Model Lemma

For any world w in any canonical model M_{AX} and any sentence A, $A \in w$ iff $M_{\mathrm{AX}}, w \models A$.

Proof: The proof is by induction on complexity. We first show that the lemma holds for sentence letters. Then we show that *if* the lemma holds for some sentences A and B, *then* it also holds for the corresponding more complex sentences $\neg A$, $A \rightarrow B$, and $\Box A$.

1. (Sentence letters.) If A is a sentence letter, then by definition 4.1, $V(w, A) = 1$ iff $A \in w$, and so by clause (a) of definition 3.2, $M_{\mathrm{AX}}, w \models A$ iff $A \in w$.
2. (Case $\neg A$.) By clause (b) of definition 3.2, $M_{\mathrm{AX}}, w \models \neg A$ iff $M_{\mathrm{AX}}, w \not\models A$. If the lemma holds for A, then $M_{\mathrm{AX}}, w \not\models A$ iff $A \notin w$. And because w is maximal, we have $A \notin w$ iff $\neg A \in w$. So $M_{\mathrm{AX}}, w \models \neg A$ iff $\neg A \in w$.
3. (Case $A \rightarrow B$.) $M_{\mathrm{AX}}, w \models A \rightarrow B$ iff either $M_{\mathrm{AX}}, w \not\models A$ or $M_{\mathrm{AX}}, w \models B$, by clause (c) of definition 3.2. Given that the lemma holds for A and B, we have $M_{\mathrm{AX}}, w \not\models A$ iff $A \notin w$, and $M_{\mathrm{AX}}, w \models B$ iff $B \in w$. So $M_{\mathrm{AX}}, w \models A \rightarrow B$ iff either $A \notin w$ or $B \in w$. And since w is maximal consistent, $A \rightarrow B \in w$ iff either $A \notin w$ or $B \in w$. So $M, w \models A \rightarrow B$ iff $A \rightarrow B \in w$.
4. (Case $\Box A$.) From left to right, suppose $\Box A \in w$. By definition 4.1, it follows that $A \in v$ for all v with wRv. Assuming that the lemma holds for A, we have $M_{\mathrm{AX}}, v \models A$ for all v with wRv. So $M_{\mathrm{AX}}, w \models \Box A$, by clause (g) of definition 3.2.

   For the converse direction, suppose $\Box A \notin w$. Let Γ^- be the set of all sentences B such that $\Box B \in w$. By observation **??**, $\Gamma^- \cup \{\neg A\}$ is AX-consistent. By

definition 4.1 and Lindenbaum's lemma, it follows that there is some $v \in W$ such that wRv and $\neg A \in v$. So $A \notin v$. Assuming that the lemma holds for A, we have $M_{\mathrm{AX}}, v \not\models A$. So $M_{\mathrm{AX}}, w \models \Box A$, by clause (g) of definition 3.2. □

The completeness of AK follows immediately from the previous two lemmas, as foreshadowed above:

Theorem: Completeness of AK

If $\models_K A$ then $\vdash_{\mathrm{AK}} A$.

Proof: We show that if $\nvdash_{\mathrm{AK}} A$ then $\not\models_K A$. Suppose $\nvdash_{\mathrm{AK}} A$. Then $\{\neg A\}$ is AK-consistent. It follows by Lindenbaum's Lemma that $\{\neg A\}$ is included in some maximal AK-consistent set S. By definition 4.1, that set is a world in M_{AK}. Since $\neg A \in S$, it follows from the Canonical Model Lemma that $M_{\mathrm{AK}}, S \models \neg A$. So $M_{\mathrm{AK}}, S \not\models A$. So A is not true at all worlds in all models: $\not\models_K A$. □

Done!

To prove completeness for axiomatic calculi stronger than AK, we need one more step. Let's do the proof for AT.

Theorem: Completeness of AT

If $\models_T A$ then $\vdash_{\mathrm{AT}} A$.

Proof: We begin as before. We'll show that if $\nvdash_{\mathrm{AT}} A$ then $\not\models_T A$. Suppose $\nvdash_{\mathrm{AT}} A$. Then $\{\neg A\}$ is AT-consistent. It follows by Lindenbaum's Lemma that $\{\neg A\}$ is included in some maximal AT-consistent set S. By definition 4.1, that set is a world in M_{AT}. Since $\neg A \in S$, it follows from the Canonical Model Lemma that $M_{\mathrm{AT}}, S \models \neg A$. So $M_{\mathrm{AK}}, S \not\models A$. This proves that there is some world in some Kripke model where A is false. But what we need to show is that there is some world in some *reflexive* Kripke model where A is false. (That's what $\not\models_T A$ means.)

Fortunately, we can show that the canonical model for AT is reflexive. By definition 4.1, a world w in a canonical model is accessible from itself iff whenever $\Box A \in w$ then $A \in w$. Since the worlds in M_{AT} are maximal AT-consistent sets of sentences, and every such set contains every instance of the **T**-schema $\Box A \rightarrow A$, there is no

world in M_{AT} that contains $\Box A$ but not A. So every world in M_{AT} has access to itself.

Above, we've shown that if $\nvdash_{\mathrm{AT}} A$ then A is false at some world in M_{AT}. Now we've shown that M_{AT} is reflexive. So if $\nvdash_{\mathrm{AT}} A$ then A is false at some world in some reflexive model, which means that $\not\models_T A$. □

Exercise 4.8

Outline the completeness proof for $\vdash_{\mathrm{AK4}}$. (You need to show that the canonical model for K4 is transitive.)

If an axiomatic calculus can prove all and only the sentences that are valid in some class of frames, the calculus is said to be **characterised** (or **determined**) by that class. Our soundness and completeness results for AK and AT show that AK is characterised by the class of all frames, while AT is characterised by the class of reflexive frames. The following table summarizes a few other results along these lines.

Axiomatic Calculus	*Characterised By*
AK	all frames
AT	reflexive frames
AD	serial frames
AK4	transitive frames
AB	reflexive and symmetric frames
AS4	reflexive and transitive frames
AS4.2	reflexive, transitive, and convergent frames
AS5	reflexive, transitive, and symmetric frames
AS5	universal frames

In the previous chapter, I gave a similar table showing, for example, that the system K contains the sentences valid on all frames, that T contains the sentences valid on all reflexive frames, and so on. But there was nothing interesting about that, because this is how I *defined* the systems K, T, etc. The present table, by contrast, shows that the syntactically defined calculi on the left are sound and complete with respect to the class of frames on the right.

It should come as no surprise that AS5 is characterised both by the class of universal frames and by the class of reflexive, transitive, and symmetric frames. We know from observation 3.5 that the exact same sentences are valid in both of these classes. In general, the fact that an axiomatic calculus is characterised by some class of frames doesn't mean that the calculus isn't also characterised by a different class of frames.

I should note that the above technique for proving completeness does not always work. In the next section, we will meet the normal axiomatic calculus AGL. AGL is characterised by a certain class of frames, but its canonical model doesn't belong to that class. Worse, the axioms of AGL are not valid on the frame of its canonical model. So the above technique can't be used to prove completeness. Other (normal) calculi are not even characterised by any class of frames. An example is the calculus AKH, which results from AK by adding the axiom schema

$$\Box(\Box A \leftrightarrow A) \rightarrow \Box A \qquad (\mathbf{H})$$

4.4 Provability logic

We've looked at some proofs about what is and isn't provable with a certain method. When we prove facts about proofs, we are doing *meta-logic*. The meta-logic proofs I gave in the last two sections were informal: I did not use a formal (natural deduction or tree or axiomatic) method to prove soundness and completeness theorems. But in principle that could be done. Of course, a proof method for propositional logic would not have been enough, since my proofs often involved quantifiers. I also used several assumptions from set theory in the proof of Lindenbaum's Lemma. So I could, in principle, have used an axiomatic calculus for (first-order) predicate logic, with some further axioms about sets. A well-known calculus of that kind is *ZFC*. (It is named after Ernst Zermelo, Abraham Fraenkel, and the Axiom of Choice). ZFC is strong enough to prove not just soundness and completeness in modal logic, but almost everything else that can be proved in any branch of maths.

An interesting feature of ZFC is that it can prove facts about provability not just in weaker axiomatic calculi like AK; it can also prove facts about provability in ZFC itself. Let $\vdash_{\text{ZFC}} A$ mean that A can be proved in ZFC. Then one can, for example, prove *in ZFC* that whenever $\vdash_{\text{ZFC}} (A \rightarrow B)$ and $\vdash_{\text{ZFC}} A$ then $\vdash_{\text{ZFC}} B$.

This brings us back to modal logic. Suppose we read the box as 'it is mathematically provable that . . . ', and we understand mathematical provability as provability in ZFC.

Then every instance of the **K**-schema

$$\Box(A \to B) \to (\Box A \to \Box B) \qquad (\mathbf{K})$$

in the language of ZFC is true. More interestingly, as I just mentioned, every such instance is provable in ZFC.

If a sentence A is provable in ZFC, then $\vdash_{\text{ZFC}} A$ is also provable in ZFC. Moreover, since ZFC is an extension of classical predicate logic, every instance of **A1**–**A3** is provable in ZFC, and **MP** is an inference rule of ZFC.

So the logic of mathematical provability is a normal modal logic: We have **A1**–**A3**, **K**, **Nec**, and **MP**. What other principles do we have? You might expect that the **T**-schema should hold:

$$\Box A \to A \qquad (\mathbf{T})$$

Surely if something is mathematically provable, then it is true. But if we're looking at what ZFC can prove about mathematical provability, it turns out that some instances of **T** are not provable. That is, while every instance of **T** is intuitively true, we have no mathematical proof of this assumption. Instead, the following shocking principle can be established (in ZFC).

$$\Box(\Box A \to A) \to \Box A \qquad (\mathbf{GL})$$

The schema is named after Kurt Gödel and Martin Löb. The calculus AGL, consisting of **A1**–**A3**, **K**, **GL**, **Nec**, and **MP**, completely captures what ZFC can prove about provability in ZFC.

To see why **GL** is shocking, consider an arbitrary mathematical falsehood – say, that 2+2=5. Obviously, 2+2=5 is not provable in ZFC. I say "obviously", but can we prove (in ZFC) that 2+2=5 is not provable? The answer is no. If we could prove $\neg\Box(2 + 2 = 5)$, then we could also prove $\Box(2 + 2 = 5) \to (2 + 2 = 5)$, which is a truth-functional consequence of $\neg\Box(2 + 2 = 5)$. And since every instance of **GL** is provable, if we could prove $\Box(2 + 2 = 5) \to (2 + 2 = 5)$ then we could also prove $2 + 2 = 5$ itself. So if we could prove (in ZFC) that we can't prove that 2+2=5, then we could also prove that 2+2=5!

What we see here is a reflection of a deep fact about the limitations of mathematical provability, known as *Gödel's second incompleteness theorem*. The theorem states that no consistent axiomatic calculus that is powerful enough to formalize elementary

mathematical reasoning (and for which there is a finite procedure for testing whether something is an axiom) can prove its own consistency. Note that an inconsistent axiomatic calculus (in classical logic) can prove everything, because everything classically follows from a contradiction. So if ZFC could prove that it *can't* prove 2+2=5, then ZFC could in effect prove its own consistency. Gödel's theorem tells us that ZFC would then be inconsistent, in which case it could prove anything, including 2+2=5.

In terms of Kripke models, the calculus AGL is characterised by to the class of Kripke models in which W is finite and R is transitive and irreflexive. (As I said earlier, the completeness proof is non-trivial, because the canonical model of AGL has infinitely many worlds and therefore does not itself belong to the relevant model class.)

The Kripke semantics for provability logic is interesting because it does not capture any intuitive meaning of the box. Intuitively, mathematical truths are true at all possible worlds, so it is hard to see how mathematical provability could be analysed in terms of truth at accessible worlds. Nonetheless, Kripke models play an important role in the study of provability: the standard method of proving that all instance of **GL** are provable in ZFC (due to Robert Solovay) involves the semantic characterisation of AGL in terms of Kripke models.

Exercise 4.9

Explain why the logic of provability does not contain all instances of **5**, given that it contains all instances of **GL**.

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 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 epistemically accessible for agent 1 but no p-world is accessible for agent 2. How many 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 operators of standard epistemic logic do not always match our everyday concept of knowledge.

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 $\mathsf{K}_i\, q$, then our Kripke semantics will guarantee that $\mathsf{K}_i\, p \wedge q$. 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 $\mathsf{K}_i\, B$.

By contrast, it is easy to imagine a scenario in which someone knows certain propositions (say, the axioms of a mathematical theory) and yet fails to know some logical consequences of these propositions. So it can be misleading to translate '$\mathsf{K}_i\, p$' as 'agent i knows that p'. More adequate translations are 'p is compatible with the information available to agent i', 'i is in a position to know p', 'given i's evidence, p must be the case', or 'i implicitly knows p'.

For the sake of brevity, we will nonetheless often paraphrase 'K' as 'knows'. But you should keep in mind that what we are modelling is a concept of implicit knowledge that is closed under logical consequence.

Exercise 5.1

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 5.2

The reason why logicians mostly focus on implicit knowledge is that the ordinary concept of knowledge is logically ill-behaved in many ways. Let K^* be an operator that applies to a sentence A iff we would intuitively say that an agent knows A (or rather, the relevant translation of A into English). 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 operator that represents mathematical

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

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

From the definition of K^+, it follows that

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

Explain why this is an intuitively unacceptable principle about knowledge.

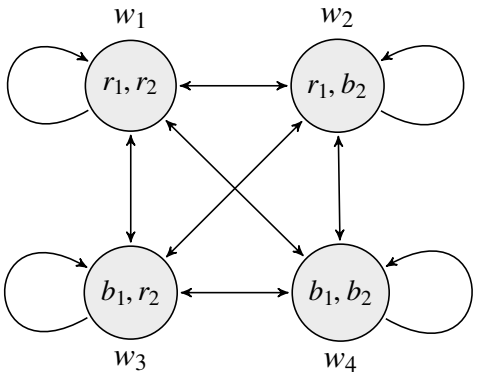
5.2 Gaining information

To get a feeling for epistemic logic, it helps to work through a few examples of how new information changes the possibilities that are open for an epistemic agent. 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.

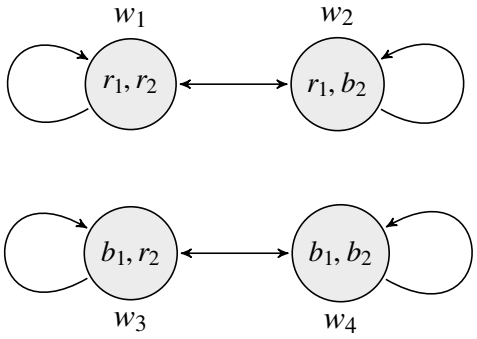
Since we are only interested in Ava's attitude towards the colour of her cards, we can take these possibilities as our possible worlds. (It usually suffices to make the worlds in our models maximally specific *with respect to the questions we're interested in*.) Here is the situation pictured as a Kripke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



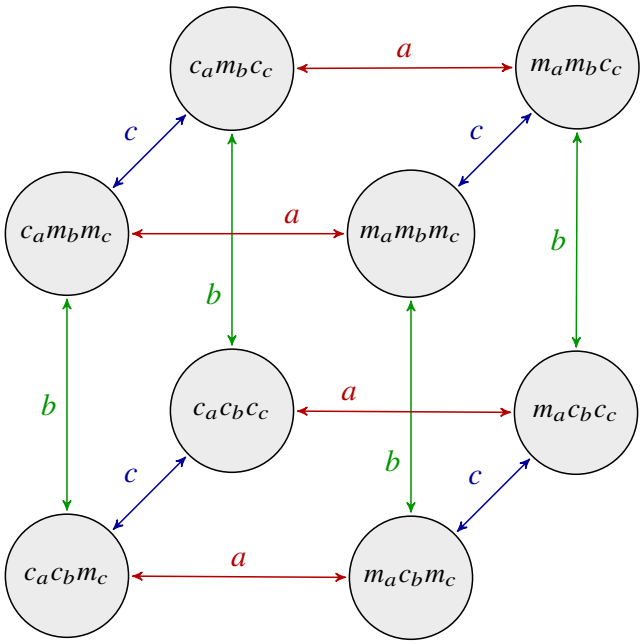
In w_1, Ava sees that the first card is red. So she can rule out worlds w_3 and w_4, where the first card is black. That's why there are no more arrows from w_1 to these worlds. But she still can't rule out world w_2, where the second card is black. If Ava then

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?

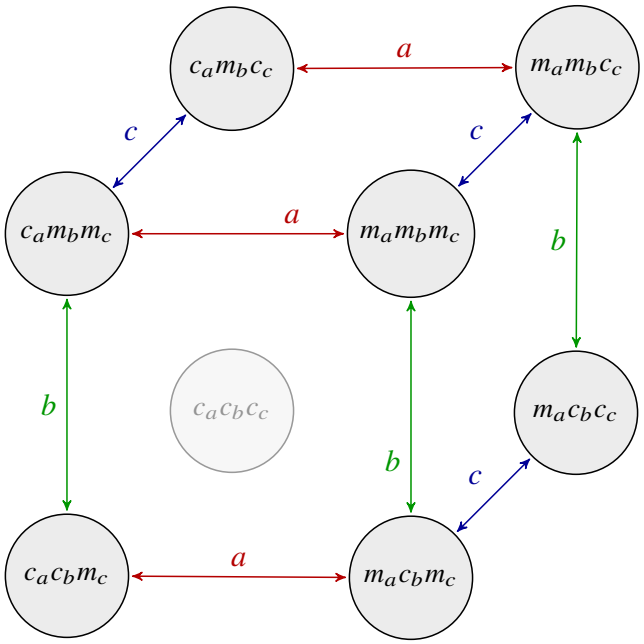
Let's draw a model. I'll call the three children Alice, Bob, and Carol, and I'll use m_a, m_b, m_c as sentence letters expressing, respectively, that Alice/Bob/Carol is muddy; c_a, c_b, c_c represent that Alice/Bob/Carol is clean. Before the mother's first announcement, there are eight possibilities.



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.

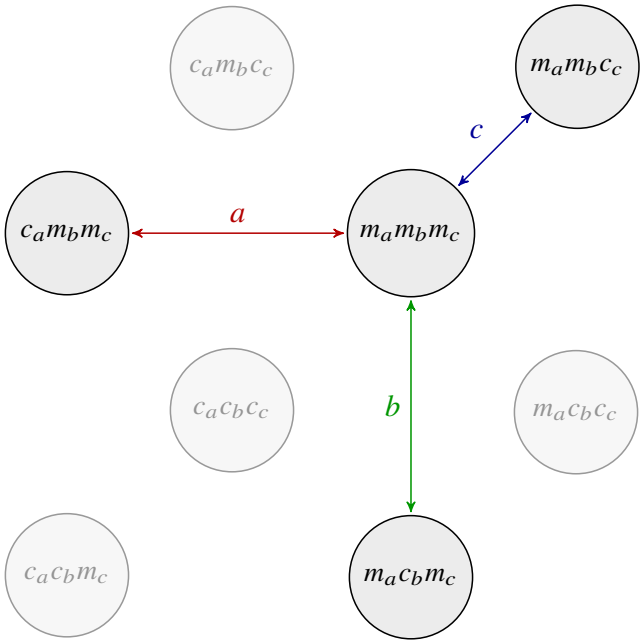
What changes through the mother's first announcement, that at least one child has mud on their face? The announcement tells *us* that we're not in the $c_a c_b c_c$ world. More importantly, it allows *each child* to rule out the $c_a c_b c_c$ world (since they all hear the announcement).



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 $c_a m_b c_c$ 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 5.3

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:

$$\mathsf{K}_i\, A \to A \tag{\textbf{T}}$$

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

$$\mathsf{K}_i\, A \to \mathsf{M}_i\, A \tag{\textbf{D}}$$

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.

$$\mathsf{K}\, A \rightarrow \mathsf{K}\, \mathsf{K}\, A. \qquad (\mathbf{4})$$

In epistemic logic, **4** is more commonly known as **positive introspection** or **the KK principle**. If you know something, does it follow that you know that you know it? Or rather, if you're in a position to know something, does it follow that you're in a position to know that you're in a position to 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, at any world w, you know 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

$$\Box(N \to (s \to p)) \wedge N. \tag{*}$$

Moreover, s indicates that s indicates that p iff

$$\Box(N \to (s \to (\Box(N \to (s \to p)) \wedge N))) \wedge N. \tag{**}$$

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

Exercise 5.4

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**.

$$\neg \mathsf{K}\, A \to \mathsf{K} \neg \mathsf{K}\, A. \tag{**5**}$$

Semantically, the **5**-schema corresponds to euclidity. Since we already have reflexivity, and reflexivity and euclidity entail symmetry (exercise 3.3), positive and negative introspection together 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 often accepted 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. It is then not implausible to say that the agent *knows* p_i just in case it stores p_i in its database. By scanning its own database, the agent can then easily find out whether or not it knows p_i. So if the agent knows p_i, then it is in a position to know that it knows p_i (**4**), and if the agent doesn't know p_i, then it is in a position to know that it doesn't know p_i (**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.

Notice that in ordinary language, saying that someone doesn't know p typically implies 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 5.5

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 false. A common view among epistemic logicians is that the logic of (implicit) knowledge lies 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 that we rejected negative introspection on the grounds that if an agent falsely believes p, then they don't know p and yet 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 argue 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:

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

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**:

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

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 5.6

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 5.7

Can you 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 5.8

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

So the multi-agent logic of knowledge is a straightforward extension of the single-agent logic of knowledge.

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 5.9

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 in a model 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:

$$\mathsf{C}_G\, A \leftrightarrow (A \wedge \mathsf{E}_G\, \mathsf{C}_G\, A) \qquad \textbf{(CK1)}$$

$$(A \wedge \mathsf{C}_G(A \rightarrow \mathsf{E}_G)) \rightarrow \mathsf{C}_G\, A \qquad \textbf{(CK2)}$$

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 boxes, '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,

$$\mathsf{B}\,A \rightarrow \neg\,\mathsf{B}\,\neg A \qquad (\mathbf{D})$$

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 belief:

$$\mathsf{B}\,A \rightarrow \mathsf{B}\,\mathsf{B}\,A \qquad (\mathbf{4})$$

$$\neg\,\mathsf{B}\,A \rightarrow \mathsf{B}\,\neg\,\mathsf{B}\,A \qquad (\mathbf{5})$$

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

Exercise 5.10

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 5.11

Show (in any way you like) that $\mathsf{B}(\mathsf{B}\,A \rightarrow A)$ 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.

$$\mathsf{K}\,A \rightarrow \mathsf{B}\,A \qquad \textbf{(KB)}$$

$$\mathsf{B}\,A \rightarrow \mathsf{K}\,\mathsf{B}\,A \qquad \textbf{(PI)}$$

$$\neg\,\mathsf{B}\,A \rightarrow \mathsf{K}\,\neg\,\mathsf{B}\,A \qquad \textbf{(NI)}$$

$$\mathsf{B}\,A \rightarrow \mathsf{B}\,\mathsf{K}\,A \qquad \textbf{(SB)}$$

KB assumes that knowledge implies belief. **PI** and **NI** strengthen the introspection principles for belief, assuming that agents always 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:

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

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 5.12

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

Exercise 5.13

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 5.14

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. For example, philosophers have often been interested not just in what we *do* know, but also in what we *can* know. Skeptical arguments suggest that we cannot know whether we have hands. 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:

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

The Knowability principle was refuted by Alonzo Church with the following neat little argument.

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 5.15

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.)

6 Deontic Logic

6.1 Permission and obligation

Deontic logic studies formal properties of obligation, permission, prohibition, and related normative concepts. The box in deontic logic is usually written 'O' (for 'obligation' or 'ought'), the diamond 'P' (for 'permission'). So if q translates 'you cook dinner', we might use $\mathsf{O}\,q$ to express that you must cook dinner, in the deontic sense of 'must': it is obligatory that you cook dinner.

We assume that obligation and permission are duals. You are not obligated to do something iff you are allowed to not do it. You are not allowed to do something iff you are obligated to not do it.

What else should we assume about the logic of obligation and permission? Does $\mathsf{O}\,A$ entail $\mathsf{P}\,A$? Does $\mathsf{P}\,A$ entail $\mathsf{P}(A \vee B)$? These are some of the questions discussed in deontic logic. At a more substantive level, we will investigate whether obligation and permission can be understood in terms of Kripke models, and if so, what these models should look like.

Before we get to all that, a few more clarifications are in order.

First, there are many kinds of norms: legal norms, moral norms, prudential norms, social norms, and so on. There may also be overarching norms that combine some or all of the others. Deontic logic is applicable to all kinds of norms, so we do not have to settle whether O expresses legal obligation or moral obligation or some other kind of obligation. However, it is important not to equivocate. If the law requires p and morality $\neg p$, we should not formalize this as $\mathsf{O}\,p \wedge \mathsf{O}\,\neg p$. It would be better to use a multi-modal language with different operators for legal and moral obligations.

Second, we need to clarify how obligations and permissions are related to agents. Intuitively, obligations and permissions vary from agent to agent. If it is your turn to cook dinner, then you are obligated to cook dinner, but I am not. To capture this agent-relativity, we could add agent subscripts to the operators, as we did in epistemic logic. We could then express our different obligations as $\mathsf{O}_1\,p \wedge \neg\,\mathsf{O}_2\,p$. But what

does the sentence letter p stand for? When I say that you are obligated to cook dinner, the object of the obligation appears to be a type of act: cooking dinner. In the language of modal propositional logic, however, O and P are sentence operators. So unless we want to say that verb phrases in English (like 'cook dinner') should be translated into sentences of $\mathfrak{L}_M$ – which is actually doable, but non-standard – we have to transform the acts an agent is obligated or permitted to perform into propositions.

For example, sentence (1) is arguably equivalent to sentence (2).

(1) You ought to cook dinner.

(2) You ought to see to it that you cook dinner.

In (2), the operator 'you ought to see to it that' attaches to a sentence, 'you cook dinner'. So we can translate (1) via (2) as $\mathsf{O}_1\, p$, where p translates 'you cook dinner', and O_1 corresponds to 'you ought to see to it that'.

The subject (you) is mentioned twice in (2), which may seem redundant. A common assumption in deontic logic is that we can drop the agent subscripts from the deontic operators, since the embedded proposition will tell us upon whom the obligation or permission falls. Informally, the idea is that (2) is equivalent to (3), with an impersonal 'ought'.

(3) It ought to be the case that you cook dinner.

The impersonal 'ought' also figures in statements like (4).

(4) Nobody ought to die of hunger.

When I say (4), I don't mean that nobody is obligated to die of hunger, nor do I mean that everybody is obligated to not die of hunger. Rather, I mean that a certain state of affairs – that nobody dies of hunger – ought to be the case. By itself, this does not impose any obligations on anyone.

On closer inspection, the equivalence between personalized 'ought' statements like (2) and impersonal 'ought' statements like (3) is questionable. Suppose Amy has promised to play with Betty. So Amy is obligated to play with Betty. But Betty is not thereby obligated to play with Amy. Betty may even have promised not to play with Amy. It is hard to express these facts in terms of impersonal oughts. If we say that it ought to be the case that Amy plays with Betty, we're missing the fact that the

obligation falls on Amy, not on Betty (who might be under a contrary obligation). So perhaps it would be better to keep the agent subscripts after all.

It can also be useful to make the 'see to it that' component in statements like (2) explicit. That Amy ought to play with Betty could then be translated as $\mathsf{O}_a\,\mathsf{STIT}\,p$, where STIT formalizes 'sees to it that'. This allows us to distinguish between the following three claims.

$\mathsf{O}_a\,\mathsf{STIT}\,\neg p$	Amy ought to see to it that she doesn't play with Betty.
$\mathsf{O}_a\,\neg\,\mathsf{STIT}\,p$	Amy ought to not see to it that she plays with Betty.
$\neg\,\mathsf{O}_a\,\mathsf{STIT}\,p$	It is not the case that Amy ought to see to it that she plays with Betty.

The STIT operator has proved useful to represent different concepts of rights and duties. In what follows, we will nonetheless stick to the simplest (and oldest) approach, without a STIT operator and agent subscripts. This approach is sufficient for many applications, but its limitations should be kept in mind.

Exercise 6.1

Let $\mathsf{F}A$ mean that A is forbidden. Can you define F in terms of O or P (or both)?

6.2 Ideal worlds

Think of a possible world as a history of events. For any such history, and any system of norms, we can ask whether the history conforms to the norms or not. Let's call a world *acceptable* (relative to some norms) if everything that happens at the world conforms to the norms. That is, a world is acceptable if it contains no violation of any relevant norm.

By definition, whatever happens at an acceptable world is permitted, in the sense that it does not violate any (relevant) norms. The converse is plausible as well: whenever something is permitted then it is the case at some acceptable world. For example, if it is permitted that Amy plays with Betty, then there should be a complete history of events in which Amy plays with Betty and no norms are violated. If there were no such history, then Amy's playing with Betty would logically entail the

violation of some norms; but if an act entails the violation of some norms, then it is hard to see how the act could be permitted relative to these norms.

So we have the following connection between permission and acceptable worlds, which amounts to a "possible-worlds analysis" of permission:

> *A* is permitted relative to some norms iff *A* is the case at some possible world that is acceptable relative to these norms.

Given the duality of permission and obligation, we also get a possible-worlds analysis of obligation:

> *A* is obligatory relative to some norms iff *A* is the case at all worlds that are acceptable relative to these norms.

For example, if it is obligatory that you cook dinner, then every history of events in which all obligations are met is a history in which you cook dinner.

As in earlier chapters, we can turn this analysis into a "model theory" for deontic logic. In logic, we are not interested in who is obligated to do what, but in whether a given deontic statement is logically valid, or whether it logically follows from other statements. Intuitively, validity means truth in every conceivable scenario under every interpretation of the descriptive vocabulary. A scenario has to settle not just the descriptive facts, but also the relevant norms, which are represented by the "acceptability relation" between possible worlds.

A package of a conceivable scenario and an interpretation of the sentence letters can therefore be compressed into a Kripke model, in which the accessibility relation is the accessibility relation.

So a standard model of (propositional) deontic logic consists of a set of "worlds" W, an "accessibility" relation on W, and an interpretation function V that assigns a truth-value to every sentence letter at every world. A world v is accessible from a world w if everything that goes on at v is permitted by the norms at w – equivalently, if everything that ought to be the case at w is the case at v. Worlds that are accessible from w in this sense are also called **ideal** relative to w.

The possible-worlds analysis of obligation and permission is reflected in definition 3.2, which settles under what conditions a sentence is true at a world in a model. Writing the box as 'O' and the diamond as P', clauses (g) and (h) of the definition

state:

$$M, w \models \mathsf{O}\, A \text{ iff } M, v \models A \text{ for all } v \text{ such that } wRv;$$
$$M, w \models \mathsf{P}\, A \text{ iff } M, v \models A \text{ for some } v \text{ such that } wRv.$$

A sentence is valid iff it is true at every world in every relevant model. If we count all Kripke models as relevant, the logic of obligation and permission will be the minimal normal modal logic K. We can get stronger logics by imposing formal constraints on the accessibility relation. Let's have a look at a few options.

If we stipulate that the deontic accessibility relation is reflexive, so that every world can see itself, then the **T**-schema for obligation becomes valid:

$$\mathsf{O}\, A \rightarrow A \qquad (\mathbf{T})$$

But there are many intuitive counterexamples to **T**. The fact that you are obligated to cook dinner does not logically entail that you cook dinner. Semantically speaking, many worlds are not ideal relative to themselves. So we should not assume reflexivity.

We might, however, impose the weaker condition of seriality – that each world can see some world. This would validate principle **D**:

$$\mathsf{O}\, A \rightarrow \mathsf{P}\, A \qquad (\mathbf{D})$$

Intuitively, **D** says that the norms are consistent: if you're obligated to do A, then you can't also be obligated to do not-A. (Remember that $\mathsf{P}\, A$ is equivalent to $\neg\, \mathsf{O}\, \neg A$.) Semantically, **D** corresponds to the assumption that there is always at least one ideal world at which all the norms are satisfied.

Without seriality, we have to allow for worlds from which no world is accessible. At such a world, all sentences of the form $\mathsf{O}\, A$ are true, and all sentences of the form $\mathsf{P}\, A$ are false. Everything is obligatory, but nothing is allowed. That makes little sense. If we use use Kripke semantics for deontic logic, we should therefore rule out inconsistent norms and accept **D** as valid.

Here it may be important to distinguish *prima facie* obligations from *actual*, or *all-things-considered* obligations. If you've promised to cook dinner, you are under a *prima facie* obligation to cook dinner. But the obligation can be overridden by intervening circumstances or contrary obligations. If your child has an accident and needs urgent medical care, the right thing to do may well be to not cook dinner

and instead bring your child to the hospital. In a sense, you are under conflicting obligations: you ought to cook dinner, and you ought to look after your child (and not cook dinner). There is no world at which you meet both of these obligations. But arguably that is not a counterexample to **D**, if we understand O as all-things-considered obligation. You are *prima facie* obligated to cook dinner, but all things considered, you should not cook dinner.

Let's return to the fact that the deontic accessibility relation is not reflexive, because many things that are not the case nonetheless ought to be the case. Some have argued that this is only true in non-ideal worlds. In an ideal world, everything that ought to be the case is the case. By this line of thought, if a world v is accessible from some world v – meaning that v is ideal relative to w – then v should be accessible from itself. This condition is sometimes called "shift reflexivity" and corresponds to the following schema **U** (for "utopia")

$$\mathsf{O}(\mathsf{O}\,A \rightarrow A) \tag{\textbf{U}}$$

In words: it ought to be the case that whatever ought to be the case is the case.

The **U** principle is entailed by an alternative way of formalizing obligation and permission that goes back to Leibniz. Let 'OK' be a propositional constant whose intended meaning is that all norms are satisfied, no obligations violated. Suppose we add this expression to the standard language $\mathfrak{L}_M$ of modal propositional logic, and we interpret the box of $\mathfrak{L}_M$ as circumstantial necessity. Arguably, $\mathsf{O}\,A$ is then definable as $\Box(\mathsf{OK} \rightarrow A)$: it ought to be that A iff, necessarily, A is the case whenever all obligations are met. It is not hard to show that if the **T**-schema is valid for the circumstantial box, then the **U**-schema is valid for O on the present definition.

Exercise 6.2

(a) Translate the **U**-schema into the Leibnizian language just proposed.
(b) Give a tree proof for the translated **U**-schema, using the T-rules for the box.

Exercise 6.3

How could we define P in terms of $\Box$ and OK?

Turning to more familiar schemas and frame conditions, what shall we say about transitivity and euclideanness, and the corresponding schemas **4** and **5**?

$$\mathsf{O}\,A \rightarrow \mathsf{O}\,\mathsf{O}\,A \tag{**4**}$$

$$\mathsf{P}\,A \rightarrow \mathsf{O}\,\mathsf{P}\,A \tag{**5**}$$

Here we face a problem. Translated back into English, it is rather unclear what these are meant to say. If something ought to be the case, ought it to be the case that it ought to be the case? If something is permitted, is it obligatory that it is permitted?

Iterations of deontic operators sound strange in ordinary language. But they have a well-defined meaning in our Kripke semantics. So let's try to figure out what **4** and **5** mean.

The validity of **4** would mean that whenever something is obligatory at a world, then it is also obligatory at all ideal alternatives to that world. Similarly, **5** would mean that if something is permissible at a world, then it's also permissible at all ideal alternatives to that world. On the background of **D**, these two assumptions would imply that for each world there is a class of ideal worlds all of which are ideal relative to each other.

To get a clearer grip on whether that is plausible, we need to clarify how obligations and permissions can vary from world to world.

One obvious sense in which norms can vary across worlds is that people subscribe to different norms at different worlds. At our world, UK traffic law requires driving on the left, and most people think it is morally wrong to torture animals for fun. At other worlds, the laws and attitudes are different.

Let v be a world at which the traffic laws require driving on the right, and at which everyone thinks it is fine to torture animals. Suppose Norman at v is torturing kittens, while driving on the right (in the UK). Is Norman doing something that's morally wrong? Is he doing something that violates the traffic laws?

The answer depends on whether we evaluate Norman's acts relative to our norms – the norms at our world – or relative to the norms at Norman's world. Both perspectives make sense, and they lead to different deontic logics.

If we hold fixed our norm when assessing events at other worlds, then it is natural to assume that the very same worlds are ideal relative to any world, so that the deontic accessibility relation is transitive and euclidean. For example, if we judge that torturing kittens is wrong, and we hold that judgement fixed when evaluating

events at other worlds, then torturing kittens is permissible at no world; so the only worlds that are accessible from any world are worlds at which no kittens are tortured. In general, the worlds accessible from any world will be all and only those worlds at which our norms are satisfied. The resulting logic of obligation and permission is KD45.

On the other hand, if we evaluate the events at other worlds relative to the local norms of the relevant worlds, then transitivity and euclidity become implausible, as does shift reflexivity. For example, consider another world u in which the law says that one should drive on the right but everyone nonetheless drives on the left. Nothing that happens at u, we may assume, violates the traffic laws of our world. So u is deontically accessible from the actual world. But if we evaluate the events at u relative to the laws at u, then much of what happens at u violates the relevant norms, so u is not deontically accessible from itself. Shift reflexivity fails. Moreover, most worlds that are deontically accessible from u will be worlds in which people drive on the right, and these worlds are not deontically accessible from our world. So transitivity also fails.

The second, "relativist", perspective seems to be more common in deontic logic. As a consequence, so-called **standard deontic logic** assumes only that the accessibility relation is serial, making the system D the complete logic of obligation and permission.

Exercise 6.4

Suppose Amy ought to either promise to help Betty or promise to help Carla. If she were to promise to help Betty, she would be obligated to help her. And if she were to promise to help Carla, she would be obligated to help Carla. So it ought to be the case that Amy is either obligated to help Betty or obligated to help Carla. In fact, Amy makes neither promise, so she is neither obligated to help Betty nor to help Carla. Explain why this casts doubt on the validity of **5**.

Exercise 6.5

Consider the **C4** schema $\bigcirc\bigcirc A \rightarrow \bigcirc A$. Show that

(a) if **U** is valid on a frame, then so is **C4**;
(b) it is not the case that if **C4** is valid on a frame, then so is **U**.

Exercise 6.6

Give either a proof or a counterexample for the following sentences, assuming the logic of obligation and permission is D.

(a) $\mathsf{O}\,p \to \mathsf{O}(p \vee q)$
(b) $\mathsf{P}(p \vee q) \to (\mathsf{P}\,p \wedge \mathsf{P}\,q)$
(c) $(\mathsf{O}\,p \wedge \mathsf{O}\,q) \to \mathsf{O}(p \wedge q)$
(d) $\mathsf{P}\,p \to \mathsf{P}(p \vee q)$
(e) $\mathsf{O}(q \wedge \neg q)$
(f) $\neg(\mathsf{O}\,p \wedge \mathsf{O}\,\neg p)$
(g) $\mathsf{O}\,\mathsf{P}\,q \vee \mathsf{P}\,\mathsf{O}\,q$

6.3 Norms and circumstances

The possible-worlds analysis I have outlined in the previous section assumes that something ought to be the case iff it is the case at all ideal worlds, where no norms are violated. Many ordinary statements about oughts and obligations do not fit this analysis.

Suppose you are walking past a drowning baby. You ought to save the baby. But are you saving the baby at every world at which no norms are violated? Clearly not. There are worlds at which the baby never fell into the pond, and others at which you are overseas and have no means to rescue the baby. These worlds need not involve any violations of norms.

The general point is that whether something ought to be the case depends not just on the norms but also on the circumstances. If you walk past a drowning baby, you ought to rescue the baby; under other circumstances (if the baby never fell into the pond), no such obligation arises. The point also applies to impersonal oughts. In worlds where an increase of greenhouse gases threatens to destabilise the climate, greenhouses gases ought to be reduced; in worlds where greenhouse gases never increased, there is no imperative for reduction.

We can account for the dependence of obligations on circumstances by changing our interpretation of the accessibility relation. Previously, we assumed that a world v is accessible from w iff all the norms of w are respected at v. On the new interpretation, we also hold fixed relevant circumstances at w. For example, if w is a world at which

you're walking past a drowning baby, then any accessible world will also be a world at which you're walking past a drowning baby.

As a first stab, we might therefore redefine deontic accessibility as follows:

> A world v is deontically accessible from a world w iff (a) the relevant circumstances at w are also the case at v, and (b) no norms from w are violated at v.

I use 'relevant circumstances' as a placeholder for whatever we hold fixed when we consider what ought to be the case. In general, we tend to hold fixed anything that can no longer be changed. If the baby has fallen into the pond at w, then there is nothing anyone can do to undo the falling; so the falling is a "relevant circumstance" that takes place at every world accessible from w. A more informative definition of relevant circumstances would have to address difficult philosophical questions about free will, among other things. Let's set these issues aside.

Clause (b) in the above definition assumes that no norms are violated at any accessible world. But if accessibility is restricted by circumstances, then this is implausible, as the relevant circumstances will often involve violations of norms.

The problem is brought ought by Arthur Prior's "Samaritan Paradox". Suppose Smith has been injured in a robbery, and Jones has the opportunity to help. We want to say that Jones ought to help the injured Smith. On the possible-worlds analysis of 'ought', this means that Jones helps the injured Smith at all worlds accessible from the actual world. It follows that Smith has been injured at all these worlds. But then all the accessible worlds contain a violation of norms. In a truly ideal world, Smith would never have been robbed and injured.

Intuitively, Jones ought to help Smith because the relevant worlds at which Jones doesn't help Smith are even *worse*, in terms of norm violations, than the worlds at which Jones helps Smith. Both kinds of worlds are bad, because Smith got robbed. But our norms don't just divide the possible worlds into good and bad; they allow for finer distinctions between bad worlds and even worse worlds. Jones ought to help Smith because that's what he does in the *best* worlds among those he can bring about, even though none of these worlds are ideal.

So here is a second pass at the revised definition of deontic accessibility.

> A world v is deontically accessible from a world w iff (a) the relevant circumstances at w are also the case at v, and (b) v is one of the

best worlds, by the norms at w, among worlds at which the relevant circumstances from w are held fixed.

In the previous section, I argued that if we hold fixed the actual norms when evaluating events at other worlds, then the same worlds will be ideal relative to all worlds and so the logic of obligation and permission will be KD45. This is no longer true on the present, revised interpretation of deontic accessibility, unless the circumstantial accessibility relation that implicitly figures in clause (a) is an equivalence relation.

For some applications, it can be useful to explicitly represent the circumstantial and deontic component of the deontic accessibility relation. To this end, we would first add a circumstantial accessibility relation to our models. Informally (and roughly), a world v is circumstantially accessible from w iff there is something one can do at w that would bring about v. Next, we need to settle which worlds in a model are better than others, relative to the norms at any given world.

Let '$u \prec_w v$' mean that world u is better than world v by the norms at w. (Roughly speaking, '$u \prec_w v$' means that w contains *fewer* violations of norms.) We assume that for any world w, the relation $\prec_w$ is asymmetric and transitive, where asymmetry means that if $u \prec_w v$ then it is not the case that $v \prec_w u$. Asymmetric and transitive relations are also known as **partial orders**.

Definition 6.1

An **deontic ordering model** consists of

- a non-empty set W (the worlds),
- a binary relation R on W (the circumstantial accessibility relation),
- for each world $w \in W$, a partial order $\prec_w$ on W (the world-relative ranking of worlds as better or worse), and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value.

Now we need to say under what conditions a sentence of the form $\mathsf{O}A$ is true at a world in a model. Informally, $\mathsf{O}A$ should be true at w iff A is true at the best worlds among those that are circumstantially accessible. Let's introduce one more piece of notation. For any set of worlds S and any partial order $\prec_w$, let $Min^{\prec_w}(S)$ be the set of

$\prec_w$-minimal members of S:

$$Min^{\prec_w}(S) =_{\text{def}} \{v \in S : \neg\exists u(u \prec_w v)\}.$$

Intuitively, $Min^{\prec_w}(S)$ contains all the worlds from S with the fewest norm violations.

Here, then, are the truth-conditions of $\mathsf{O}\,A$ and $\mathsf{P}\,A$ in deontic ordering models M:

$$M, w \models \mathsf{O}\,A \text{ iff } M, v \models A \text{ for all } v \in Min^{\prec_w}(\{u : wRv\})$$
$$M, w \models \mathsf{P}\,A \text{ iff } M, v \models A \text{ for some } v \in Min^{\prec_w}(\{u : wRv\})$$

Don't be afraid of all the symbols. This is just a formal way of saying that $\mathsf{O}\,A$ is true at w iff A is true at the best worlds (by the norms at w) among the worlds circumstantially accessible from w.

If we want the **D**-schema to be valid, we have to assume that for any world w, there is always at least one best world among the circumstantially accessible worlds, so that $Min^{\prec_w}(\{u : wRv\})$ is never the empty set. Let's make this assumption.

Ordering models help to clarify how the logic of obligation and permission depends on formal properties of circumstantial accessibility and the deontic orderings. They also help with a common problem that often arises when we try to formalize statements containing modal operators and if-clauses.

The following puzzle is due to Roderick Chisholm.

(1) Jones ought to help his neighbours.

(2) If Jones is going to help his neighbours, then he ought to tell them he's coming.

(3) If Jones isn't going to help his neighbours, then he ought to not tell them he's coming.

(4) Jones won't help his neighbours.

It is easy to imagine a scenario in which all of (1)–(4) are true. But if we want to translate these sentences into the language of propositional deontic logic, we run into problems. (1) and (4) are simple:

(1′) $\mathsf{O}\,h$

(4′) $\neg h$

For (2), we might try (2N):

(2N) $h \to \mathsf{O}\, t$

However, remember that a material conditional $A \to B$ is true whenever the antecedent A is false. So (2N) is logically entailed by (4′). Yet intuitively, (2) is not entailed by (4): it is easy to imagine a scenario in which (4) is true but (2) false.

So it might be better to translate of (2) as (2W):

(2W) $\mathsf{O}(h \to t)$

Here the operator O is said to have **wide scope** because it applies to the entire conditional $h \to t$. In (2N), the operator has **narrow scope** because it only applies to the consequent t.

For (3), we also have a choice between a wide scope translation and a narrow scope translation.

(3N) $\neg h \to \mathsf{O}\, \neg t$

(3W) $\mathsf{O}(\neg h \to \neg t)$

(3W) is logically entailed by (1′). For note that $\neg h \to \neg t$ is true whenever h is true. So if h is true at all deontically accessible worlds, as (1′) says, then $\neg h \to \neg t$ is also true at all deontically accessible worlds. Yet intuitively, (1) does not entail (3). Again, one can imagine scenarios in which (1) is true but (3) false.

So perhaps we should use a mixed approach: translate (2) as (2W) and (3) as (3N). But then the four assumptions come out logically inconsistent!

Exercise 6.7

Give a tree proof with the D rules to show that (1′), (2W), (3N), and (4′) are jointly inconsistent. That is, start the tree with the four assumptions and show that all branches close.

There are other problems with both the narrow-scope and the wide-scope translation of 'if . . . ought . . . ' sentences. For example, intuitively (5) does not entail (6).

(5) If you have promised to call your parents, then you should call them.

(6) If you have promised to call your parents and you know that someone has attached a bomb to your parents's phone that will go off if you call, then you should call them.

But $p \rightarrow \mathsf{O}\, r$ logically entails $(p \wedge q) \rightarrow \mathsf{O}\, r$. (Do a tree proof if you can't see this!) Moreover, in the minimal modal logic K, $\mathsf{O}(p \rightarrow r)$ entails $\mathsf{O}((p \wedge q) \rightarrow r)$. So both the narrow-scope and the wide-scope translation would make the inference from (5) to (6) valid.

A popular response to these problems is to introduced a new, binary operator for **conditional obligation**. The operator is often written '$\mathsf{O}(\cdot/\cdot)$', with a slash separating the two argument places. Intuitively, $\mathsf{O}(p/q)$ means that if condition q is satisfied then p ought to be the case. The precise semantics of $\mathsf{O}(\cdot/\cdot)$ is a matter of debate; I will sketch one approach, drawing on ideas from Bengt Hansson in logic and Angelika Kratzer in linguistics.

Let's have another look at sentence (3) in Chisholm's puzzle.

(3) If Jones isn't going to help his neighbours, then he ought to not tell them he's coming.

At all ideal worlds, we may assume, Jones helps his neighbours. So (3) doesn't talk about what happens at ideal worlds. It doesn't even talk about what happens at the best worlds among those Jones can still bring about, for Jones *could* help his neighbours. Rather, (3) seems to say that among the non-ideal worlds at which Jones doesn't help his neighbours, the best worlds are worlds at which he doesn't tell them that he will come.

This suggests that the if-clause in 'if ... ought ...' sentences can restrict the worlds over which the modal operator quantifies, so that 'if p then ought q' is true iff q is true at the best of the circumstantially accessible worlds *at which p is true*.

With deontic ordering models, we can easily turn this idea into a formal semantics for the $\mathsf{O}(\cdot/\cdot)$ operator:

$$M, w \models \mathsf{O}(B/A) \text{ iff } M, v \models B \text{ for all } v \in Min^{\prec_w}(\{u : wRv \text{ and } M, v \models A\})$$

Intuitively, when we evaluate $O(B/A)$, we simply add the assumption A to the circumstances that are held fixed.

The present semantics for conditional obligation validates the principle of **deontic detachment**: $\mathsf{O}\, A$ and $\mathsf{O}(B/A)$ together entail $\mathsf{O}(B)$. For suppose A is true at the best of the (circumstantially) accessible worlds, and B is true at the best of the accessible worlds at which A is true. Then evidently B is true at the best of the accessible worlds. On the other hand, we do not have **factual detachment**: A and $\mathsf{O}(B/A)$ does not entail $\mathsf{O}(B)$.

The invalidity of factual detachment is illustrated by the "gentle murder puzzle". Suppose John is determined to kill his grandmother. *If he will go ahead and kill her, he ought to do so gently*. Can we conclude that John ought to gently kill his grandmother? Arguably not. He shouldn't kill her at all! We have k and $\mathsf{O}(g/k)$, but not $\mathsf{O}(g)$.

Exercise 6.8

The dual of conditional obligation is conditional permission.

(a) Explain why 'if you have a disability, you can park in front of the entrance' is not adequately translated as either $d \rightarrow \mathsf{P}(p)$ or $\mathsf{P}(d \rightarrow p)$.

(b) Outline a semantics for $\mathsf{P}(B/A)$ that parallels the semantics I have outlined for $\mathsf{O}(B/A)$.

6.4 Further challenges

The formalization of conditional obligations is one challenge for standard deontic logic, but it is not the only one. I will mention three others.

First, we already saw that standard deontic logic does not allow for conflicting obligations. Suppose you have promised your family to be home for dinner and your friends to join them at the pub. You are under conflicting *prima facie* obligations, and it is not clear that one of them overrides the other. Legal systems can also contain contradictory rules, without any higher-level rules for how to resolve such contradictions.

We can, of course, drop principle **D**. But even in the minimal normal logic K, $\mathsf{O}\,p$ and $\mathsf{O}\,\neg p$ entail $\mathsf{O}\,A$, for any sentence A. Intuitively, however, the fact that you have given incompatible promises does not entail that you are obligated to, say, kill the Prime Minister.

A second family of problems arises from the fact that even in the minimal modal logic K, O is closed under logical consequence: if $\mathsf{O}(A)$ is true and A entails B, then $\mathsf{O}(B)$ is true. Since logical truths are logically entailed by everything, it follows that all logical truths come out as obligatory: it ought to be that 2+2=4, or that it either rains or doesn't rain. This is easy to see semantically. Any logical truth is true at all worlds; so it is also true at all deontically accessible worlds.

In response, one might argue that the relevant statements sound wrong not because they are false, but because their utterance would violate a **pragmatic** norm of cooperative communication. A basic norm of pragmatics is that utterances should make a helpful contribution to the relevant conversation. In a normal conversational context, it would be pointless to say that something ought (or ought not) to be the case if it is logically guaranteed to be the case anyway. An utterance of 'it ought to be that p' is pragmatically appropriate only if p could be false. This might explain why it sounds wrong to say that it ought to either rain or not rain.

Note also that by duality, $\neg \mathsf{O}(p \vee \neg p)$ entails $\mathsf{P} \neg(p \vee \neg p)$. So if we deny that it ought to either rain or not rain, and we accept the duality of obligation and permission, we have to say that it is allowed that it neither rains nor doesn't rain. That sounds even worse.

The problem of closure under entailment has special bite when obligation statements are restricted by circumstances. Return to the Samaritan puzzle. Suppose Smith is bleeding, and Jones ought to stop the blood flow. It is logically impossible to stop a blood flow if no blood is flowing. In standard deontic logic, the claim that Jones ought to stop Smith's blood flow therefore entails that Smith ought to be bleeding.

Here, too, one might appeal to a pragmatic explanation. When we say that Jones ought to stop the blood flow, we take for granted that Smith is bleeding. We are interested in what should be done *given* the state in which Jones found Smith. Worlds where Smith isn't injured are set aside; they are not circumstantially accessible. But circumstantial acceptability can shift with conversational context. The claim that Smith ought to be bleeding is pointless if we hold fixed Smith's state of injury. So when we evaluate this claim, we naturally assume that the relevant circumstantial accessibility relation does not hold fixed Smith's injuries. Intuitively, we are no longer considering what should be done given the state in which Jones found Smith, but whether that state itself should have obtained. So worlds in which the state doesn't obtain become circumstantially accessible.

A third family of problems arises from disjunctive statements of permission and obligation. Consider (1).

(1) You ought to either mail the letter or burn it.

Intuitively, (1) suggests that both mailing the letter and burning it are permitted. In standard deontic logic, however, $\mathsf{O}(A \vee B)$ does not entail $\mathsf{P}\, A \wedge \mathsf{P}\, B$. (This puzzle was first noticed by Alf Ross and is known as "Ross's Paradox".)

A similar puzzle arises for permissions. (This one is known as the "Paradox of Free Choice".)

(2) You may have beer or wine.

Intuitively, (2) implies that beer and wine are both permitted. But in standard deontic logic, $\mathsf{P}(A \vee B)$ does not entail $\mathsf{P}\,A \wedge \mathsf{P}\,B$.

Can't we simply add the missing principles?

$$\mathsf{O}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B) \qquad (\mathbf{R})$$

$$\mathsf{P}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B) \qquad (\mathbf{FC})$$

No. Both of these have unacceptable consequences when added to the minimal modal logic K. With the help of **R**, we could show that $\mathsf{O}\,A$ entails $\mathsf{P}\,B$: $\mathsf{O}\,A$ entails $\mathsf{O}(A \vee B)$, which by **R** entails $\mathsf{P}\,B$. But clearly 'you ought to mail the letter' does not entail 'you may burn the letter'.

Similarly for **FC**. In the minimal modal logic K, $\mathsf{P}\,A$ entails $\mathsf{P}(A \vee B)$; by **FC**, $\mathsf{P}(A \vee B)$ entails $\mathsf{P}\,B$. But clearly 'you may have beer' does not entail 'you may have wine'.

Exercise 6.9

Show that Ross's Paradox and the Paradox of Free Choice also arise for epistemic modals.

6.5 Neighbourhood semantics

I have outlined several moves one could make to defend a Kripke-type semantics for deontic logic, despite a number of problems the semantics seems to create. Not everyone is convinced by these moves. Some hold that the true logic of obligation and permission is *non-normal*, in the technical sense that logical validity cannot be defined as truth at all worlds in some class of Kripke frames.

We then need a new semantics for our language: a new conception of a model, and a new definition of truth relative to a model. One popular alternative to Kripke semantic is **neighbourhood semantics** (also known as **Scott-Montague semantics**, after its inventors Dana Scott and Richard Montague).

Models in neighbourhood semantics still involve possible worlds, and validity is still defined as truth at all worlds in all (relevant) models. But the box and the diamond are no longer interpreted as quantifiers over accessible worlds. Instead, we simply assume that at every world, some propositions are "necessary" and others are not. $\Box A$ is true at a world if A expresses one of the necessary propositions at that world.

Formally, the accessibility relation in Kripke models is replaced by a **neighbourhood function** N that associates each world in a model with the propositions that are necessary relative to w. Propositions are identified with sets of possible worlds. So $N(w)$ is a set of sets of worlds. Each set of world in $N(w)$ is necessary at w.

Definition 6.2

A **neighbourhood model** consists of

- a non-empty set W,
- a function N that assigns to each member of W a set of subsets of W, and
- a function V that assigns to each sentence letter of $\mathfrak{L}_M$ and each member of W a truth-value.

The interpretation of non-modal sentences at neighbourhood models is just like in Kripke semantics (definition 3.2). To state the semantics for modal sentences, let $[A]^M$ be the set of worlds in model M at which A is true. This is our model proxy for the proposition expressed by A. Then:

$$M, w \models \Box A \text{ iff } [A]^M \text{ is in } N(w).$$
$$M, w \models \Diamond A \text{ iff } [\neg A]^M \text{ is not in } N(w).$$

Intuitively, the clause for the box says that $\Box A$ is true at w iff the proposition expressed by A is one of those that are necessary at w. The clause for the diamond ensures that the box and the diamond are duals.

In neighbourhood semantics, the modal operators are not closed under logical consequence. For example, the neighbourhood function N can easily make p necessary at a world without making $p \vee q$ necessary, even thought p entails $p \vee q$.

Exercise 6.10

What formal condition on the neighbourhood function would ensure that $\Box$ is closed under logical consequence?

If we interpret O and P as the box and the diamond in neighbourhood semantics, we can therefore say that Jones ought to tend to Smith's injuries even thought it is not the case that Smith ought to be injured.

We can also allow for conflicting obligations. If the laws at w require both p and $\neg p$, we simply have $[p]^M \in N(w)$ and $[\neg p]^M \in N(w)$. It longer follows that any proposition whatsoever is obligatory.

We may also escape the problems from section 6.3 that led us to introduce a primitive conditional obligation operator. I argued that the wide-scope translation $\mathsf{O}(A \to B)$ of conditional obligation sentences is problematic because $\mathsf{O}(A \to B)$ is entailed by $\mathsf{O}(\neg A)$ as well as by $\mathsf{O}(B)$. In neighbourhood semantics, both of these entailments fail.

Bare neighbourhood semantics determines a very weak logic. By imposing conditions on the neighbourhood function N, we can get a stronger logic, with more validities.

For example, suppose we want to maintain that if something is logically guaranteed to be true, then it can't be forbidden. Equivalently, any logically necessary truth should be permissible. By the neighbourhood semantics for P, A is permissible at a world w in a model M iff $[\neg A]^M$ is not in $N(w)$. If A is a logical truth, then A is true at all worlds; so $\neg A$ is true at no worlds, and $[\neg A]^M$ is the empty set. So if we want logical truths to be permissible, we have to stipulate that $N(w)$ never contains the empty set.

In Kripke semantics, the assumption that logically necessary truths are permissible is equivalent to the assumption that $\mathsf{O}\,A \to \mathsf{P}\,A$ is valid. Both assumptions require seriality of the accessibility relation. In neighbourhood semantics, we can distinguish between the two assumptions. While the permissibility of logical truths requires that $N(w)$ never contains the empty set, the validity of $\mathsf{O}\,A \to \mathsf{P}\,A$ requires that $N(w)$ never contains contradictory propositions $[A]^M$ and $[\neg A]^M$.

If we assume that the neighbourhood function is closed under intersection, in the sense that whenever two sets X and Y are in $N(w)$, then so is their intersection $X \cap Y$, then $(\Box A \wedge \Box B) \to \Box(A \wedge B)$ becomes valid. If we also require the converse,

that whenever $X \cap Y \in N(w)$ then $X \in N(w)$ and $Y \in N(w)$, and in addition that $W \in N(w)$, we get back the minimal normal logic K.

Exercise 6.11

Can you find a condition on the neighbourhood function that renders the **T**-schema valid?

Neighbourhood semantics is useful for many applications where we want a modal logic weaker than the minimal modal logic K. For some purposes, however, even the minimal logic of neighbourhood semantics is too strong. For example, return to the intuitive "Free Choice" principle from the previous section:

$$\mathsf{P}(A \vee B) \rightarrow (\mathsf{P}\,A \wedge \mathsf{P}\,B) \qquad (\mathbf{FC})$$

We have seen that this principle is untenable in Kripke semantics. As it turns out, it is still untenable in neighbourhood semantics.

To see why, note first that whenever two sentences A and B are logically equivalent, then in neighbourhood semantics $\mathsf{P}\,A$ and $\mathsf{P}\,B$ are also equivalent. The reason is that the modal operators in neighbourhood semantics operate on the set of worlds at which the embedded sentence is true. If A and B are logically equivalent, then in any model M, the set $[A]^M$ is the same set as $[B]^M$, and so $[A]^M \in N(w)$ iff $[B]^M \in N(w)$. Likewise, $[\neg A]^M \in N(w)$ iff $[\neg B]^M \in N(w)$.

Now any sentence A is logically equivalent to $(A \wedge B) \vee (A \wedge \neg B)$, for any B. In neighbourhood semantics, $\mathsf{P}\,A$ therefore entails $\mathsf{P}((A \wedge B) \vee (A \wedge \neg B))$. By **FC**, $\mathsf{P}((A \wedge B) \vee (A \wedge \neg B))$ entails $\mathsf{P}(A \wedge B)$. So by **FC**, we could still reason from 'you may have a cookie' to 'you may have a cookie and burn down the house'.

Exercise 6.12

Rational beliefs come in degrees, which are often assumed to satisfy the formal rules of probability. Suppose we say that someone believes A iff their degree of belief in A is above a certain threshold, say, 0.9. Explain why we can't give a Kripke semantics for this concept of belief. (Although one can give a neighbourhood semantics.) *Hint*: Consider the relationship between believing a conjunction and believing the conjuncts.

7 Temporal Logic

7.1 Reasoning about time

It is currently raining in Edinburgh. But it wasn't raining yesterday, and perhaps it won't rain tomorrow. Let's introduce some operators to formalize reasoning about the unfolding of events through time.

If r expresses that it is raining, we will use $\mathsf{F}\,r$ to express that is will be raining, at some point in the future. We will use $\mathsf{P}\,r$ to express that it has been raining, at some point in the past. In general:

> $\mathsf{F}\,A$ is true at a time t iff A is true at some time after t.
> $\mathsf{P}\,A$ is true at a time t iff A is true at some time before t.

The temporal operators F and P can be nested. For example, we can use $\mathsf{F}\,\mathsf{P}\,r$ to express that at some point it will have rained. $\mathsf{P}\,\mathsf{F}\,r$ means that it was once going to rain, $\mathsf{P}\,\mathsf{P}\,r$ that there was a time before which it rained, and $\mathsf{F}\,\mathsf{F}\,r$ that there will come a time after which it will rain.

Unlike $\Box$ and $\Diamond$, F and P are not duals of each other: $\neg\,\mathsf{P}\,A$ is not equivalent to $\mathsf{F}\,\neg A$, and $\neg\,\mathsf{F}\,A$ is not equivalent to $\mathsf{P}\,\neg A$. But it is useful to have duals of F and P. So we introduce two more operators. G will be the dual of F, and H be the dual of P.

Intuitively, $\mathsf{G}\,A$ means that A is always *g*oing to be the case. (Hence the symbol 'G'.) For example, if it is not the case that at some point in the future it will not rain ($\neg\,\mathsf{F}\,\neg r$), then it is always going to be the case that it will rain ($\mathsf{G}\,r$). Similarly, $\mathsf{H}\,A$ means that A *h*as always been the case. If it is not the case that at some point in the past it was not raining ($\neg\,\mathsf{P}\,\neg r$), then it has always been raining ($\mathsf{H}\,r$).

We can state the truth-conditions of $\mathsf{G}\,A$ and $\mathsf{H}\,A$ in parallel to the above truth-conditions for $\mathsf{F}\,A$ and $\mathsf{P}\,A$:

> $\mathsf{G}\,A$ is true at a time t iff A is true at all times after t.
> $\mathsf{H}\,A$ is true at a time t iff A is true at all times before t.

The language of standard propositional logic, extended by the four operators $\mathsf{F}, \mathsf{P}, \mathsf{G}, \mathsf{H}$ is known as the **language of basic temporal logic**. I'll sometimes call it $\mathfrak{L}_t$ for brevity.

Exercise 7.1

Translate the following sentences into the language of basic temporal logic.

(a) It has never been warm.
(b) It will be cold.
(c) It will not have been cold.
(d) At some point, it will be warm or it will have been warm.
(e) If you haven't studied, you won't pass the exam.
(f) I was having tea when the door bell rang.

7.2 Temporal models

In the previous section, I have informally specified the truth-conditions of $\mathsf{F}\,A$, $\mathsf{P}\,A$, $\mathsf{G}\,A$, and $\mathsf{H}\,A$ by quantifying over times: $\mathsf{F}\,A$ is true at a time iff A is true at all later times, and so on. This suggests that the truth-value of every $\mathfrak{L}_T$-sentence at any time in any conceivable scenario is fixed by the truth-value of the sentence letters at each time. Once you know at which times p, q, r, etc. are true, you can figure out the truth-value of every $\mathfrak{L}_T$-sentence at any time. For example, if p is true at some time after t, then $\mathsf{F}\,p$ is true at t, $\mathsf{G}\,\neg p$ is false at t, and $\mathsf{G}\,\mathsf{P}\,\mathsf{F}\,p$ is true at t.

For the purposes of temporal logic, we may therefore represent a scenario and an interpretation of the non-logical vocabulary by a structure that settles (a) what times there are, (b) which times come before or after which others, and (c) which sentence letters are true at which times.

Definition 7.1: Temporal Model

A **temporal model** consists of

- a non-empty set T (of "times"),
- a binary relation $<$ on T (the **precedence relation**),
- a function V that assigns to each sentence letter of $\mathfrak{L}_T$ and each member of T a truth-value (1 or 0).

I use '$M, t \models A$' as a short-hand notation to express that the sentence A is true at time t in model M. The following definition fixes the truth-value of every $\mathfrak{L}_T$-sentence at every time in every model.

Definition 7.2: Standard Temporal Semantics

If $M = \langle T, <, V \rangle$ is a temporal model, t is a member of T, ρ is any sentence letter, and A, B are any $\mathfrak{L}_T$-sentences, then

(a)	$M, t \models \rho$	iff $V(\rho, t) = 1$.
(b)	$M, t \models \neg A$	iff $M, t \not\models A$.
(c)	$M, t \models A \wedge B$	iff $M, t \models A$ and $M, t \models B$.
(d)	$M, t \models A \vee B$	iff $M, t \models A$ or $M, t \models B$.
(e)	$M, t \models A \rightarrow B$	iff $M, t \models B$ or $M, t \not\models A$.
(f)	$M, t \models A \leftrightarrow B$	iff $M, t \models (A \rightarrow B)$ and $M, t \models (B \rightarrow A)$.
(g)	$M, t \models \mathsf{F}\, A$	iff $M, s \models A$ for some $s \in T$ such that $t < s$.
(h)	$M, t \models \mathsf{G}\, A$	iff $M, s \models A$ for all $s \in T$ such that $t < s$.
(i)	$M, t \models \mathsf{P}\, A$	iff $M, s \models A$ for some $s \in T$ such that $s < t$.
(j)	$M, t \models \mathsf{H}\, A$	iff $M, s \models A$ for all $s \in T$ such that $s < t$.

Clause (a) says that a sentence letter is true at a time in a model iff the model's interpretation function of specifies that the sentence letter is true at that time. Clauses (b)–(f) say that the truth-functional connectives have their normal truth-table meaning at each time. Clauses (g)–(j) formalize the truth-conditions for temporal sentences from the previous section.

We can now formalize the intuitive notions of logical truth and logical consequence for the temporal language. We'll say that an $\mathfrak{L}_T$-sentence is **valid** iff it is true at every time in every (suitable) temporal model; an $\mathfrak{L}_T$-sentence B is a **logical consequence** of sentences $A_1, \ldots, A_n$ iff B is true at every time in every (suitable) model at which $A_1, \ldots, A_n$ are all true. I say "suitable" because we will want to put some constraints on the precedence relation $<$. More on that in a moment.

All this should remind you of our Kripke semantics for $\mathfrak{L}_M$ in chapter 3. In fact, temporal models *are* Kripke models, as defined on page 46. I have merely relabelled the set 'W' as 'T', and the relation 'R' as '$<$'.

Definition 7.2 resembles definition 3.2 from page 46, except that we have two box-like operators G and H, and two diamond-like operators F and P. The language of basic temporal logic is *bi-modal*, with forward-looking operators (F and G) and backward-looking operators (P and H) which are not definable in terms of each other. However, unlike ordinary models for multi-modal languages (definition 5.1), temporal models have only a single accessibility relation. That's because the accessibility relation for P and H is definable from the accessibility relation for F and G: a time s is earlier than a time t iff t is later than s.

Let's look at an example of a temporal model. For the set of "times" T, we choose the set of natural numbers 0,1,2, etc. Let's say that the precedence relation $<$ holds between t and s iff t is smaller than s. So $0 < 1$ and $1 < 25$, for example. (Note that we could just as well have stipulated that $<$ holds between t and s iff t is greater than s; we would then have $1 < 0$ and $25 < 1$. In temporal logic, the symbol '$<$' means 'earlier than', not 'smaller than'.) Finally, let's say that the interpretation function assigns the truth-value 1 (True) to p at a time t iff t is an even number.

Let's call this model M. By definition 7.2, we can figure out the following facts, among others.

- $M, 0 \models p$ (because $V(p, 0) = 1$);
- $M, 0 \models \mathsf{F}\, p$ (because $V(p, 2) = 1$ and $0 < 2$);
- $M, 0 \models \mathsf{G}\,\mathsf{F}\, p$ (because for every number there is a greater number that is even);
- $M, 0 \models \neg\, \mathsf{F}\,\mathsf{G}\, p$ (because there is no number for which all greater numbers are even).

Exercise 7.2

Consider another model M. As before, T is the set of natural numbers $\{0, 1, 2, \ldots\}$, and $t < s$ iff t is smaller than s. This time, $V(p, t) = 1$ iff $p < 10$. Which of the following statements are true?

(a) $M, 0 \models \mathsf{F}\, p \wedge \mathsf{F}\, \neg p$
(b) $M, 0 \models \mathsf{G}\, \neg p$
(c) $M, 0 \models \mathsf{F}\,\mathsf{G}\, \neg p$
(d) $M, 0 \models \mathsf{G}\,\mathsf{F}\, p$
(e) $M, 0 \models \mathsf{G}(\mathsf{F}\, p \rightarrow \mathsf{F}\,\mathsf{F}\, p)$
(f) $M, 0 \models \mathsf{F}\,\mathsf{H}\, p$

(g) $M, 0 \models \neg \mathsf{P}(p \vee \neg p)$
(h) $M, 0 \models \mathsf{H}\, p$

Real times are, of course, not numbers. When I say that 'it is raining' is true now, I don't mean that the sentence is true at a particular number. It isn't obvious what kinds of things times are. Fortunately, this doesn't matter for us, just as the nature of possible worlds doesn't matter for the logic of possibility and necessity. As long as the formal structure of the times in a scenario matches the structure of the natural numbers, it does no harm to use numbers as times in a model of the scenario.

The formal structure of time in a temporal model is captured by the relevant **frame**: the pair $\langle T, < \rangle$ of the set of times and the precedence relation. Frames in temporal logic are also called **flows of time**. Different applications of temporal logic often come with different assumptions about the flow of time.

In computer science, for example, the "times" T are often understood as possible states of a computational process; the precedence relation holds between states t and s if the relevant process can lead from state t to state s. If the process is indeterministic so that a given state can have different successors, the relevant flow of time will involve forks towards the future: we can have different "times" s and r such that $t < s$ and $t < r$ but neither $s < r$ nor $r < s$. Here the precedence relation cannot be modelled by the less-than relation on the natural numbers, because the structure of the less-than relation does not include forks.

In another application, we may be interested in how the weather changes from day to day. Here we might identify the relevant times with days and the precedence relation with the earlier-relation between days, even though intuitively a day is not a single time, but an interval comprising many times. For this application, the natural numbers would perhaps have the right formal structure.

For other applications, we may want to assume that time is **dense**, meaning that whenever $t < s$ then there is another point of time lying in between t and s. This assumption is common in physics. The natural numbers, by contrast, have a **discrete** structure. For example, there is no number in between 2 and 3. For dense models, we could use real or rational numbers (fractions) instead of natural numbers.

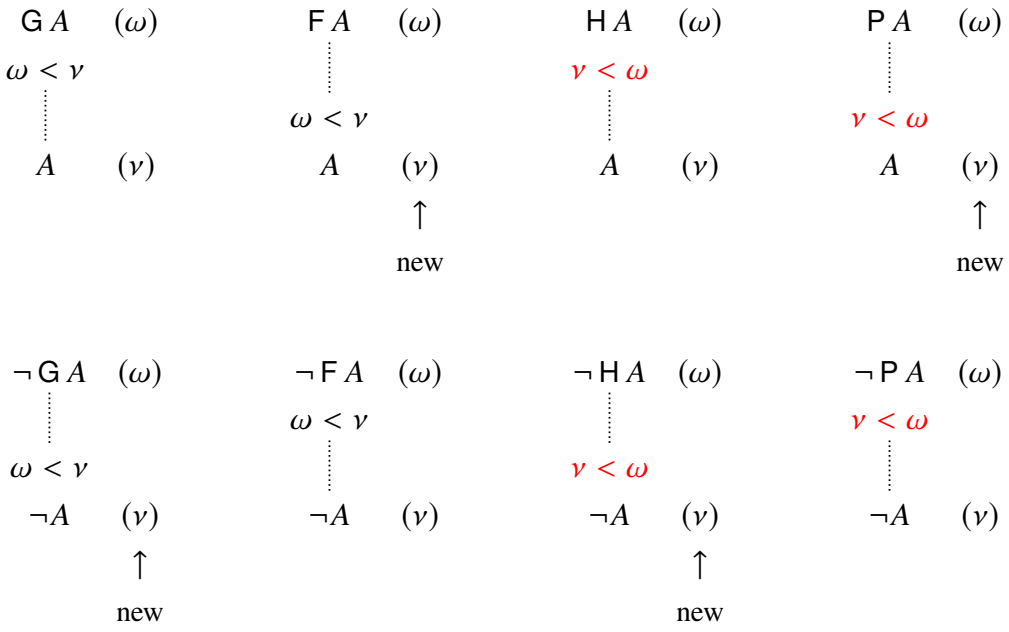
If we want to take seriously what physics tells us about time, it is not enough to assume that time is dense. We also need to reconceptualize the set T. According to the theory of special relativity, whether a point in time is earlier or later than another

is relative to a spatial frame of reference. An adequate model of relativistic time must therefore include a representation of space. In these **spacetime models** (or *Minkowski models*), the set T consists of spacetime points $\langle x_1, x_2, x_3, t \rangle$ with three spatial and one temporal coordinate; $(x_1, x_2, x_3, t) < (y_1, y_2, y_3, s)$ holds iff the second point can be reached from the first without travelling faster than the speed of light.

7.3 Logics of time

Since temporal models are just Kripke models, the standard proof systems for the minimal modal logic K also apply to temporal logic. They apply twice over, once for the forward-looking operators F and G, and once for the backward-looking P and H.

For example, the tree rules for minimal temporal logic are just the K-rules, but with the accessibility relation reversed for H and P.



In the axiomatic approach, we have two versions of the **K** schema, one for the forward-looking box G and one for the backward-looking box H:

$$\mathsf{G}(A \to B) \to (\mathsf{G}\,A \to \mathsf{G}\,B) \qquad (\mathbf{FK})$$

$$\mathsf{H}(A \to B) \to (\mathsf{H}\,A \to \mathsf{H}\,B) \qquad \textbf{(BK)}$$

Similarly, we have two versions of the Necessitation rule:

If A occurs in a proof, $\mathsf{G}\,A$ may be appended. **(FNec)**

If A occurs in a proof, $\mathsf{H}\,A$ may be appended. **(BNec)**

In addition, we need two interaction principles, reflecting the fact that the accessibility relation for F and G is the inverse of the accessibility relation for P and H.

$$A \to \mathsf{G}\,\mathsf{P}\,A \qquad \textbf{(Con1)}$$

$$A \to \mathsf{H}\,\mathsf{F}\,A \qquad \textbf{(Con2)}$$

These axioms and rules, added to those of classical propositional logic, define an axiomatic calculus that is sound and complete with respect to the class of all temporal models. (Completeness is easily proved with the canonical model technique.)

Exercise 7.3

Show that **Con1** and **Con2** are valid in the class of all temporal models.

Exercise 7.4

Show that **Con1** and **Con2** can be derived by the tree rules for minimal temporal logic.

The set of sentences that are valid at all times in all temporal models is sometimes called system K_t, because it is the temporal analogue of system K.

For most applications, this system is too weak. Not every model in the sense of definition 7.1 is an adequate model of time.

For example, definition 7.1 allows for cases in which $t < s$ and $s < r$ without $t < r$. But if a time t is earlier than s, and s is earlier than r, then surely t must be earlier than r. For almost every application of temporal logic, we will therefore want to assume that the precedence relation is transitive. This makes the **4**-schema for G valid.

$$\mathsf{G}\,A \to \mathsf{G}\,\mathsf{G}\,A \qquad \textbf{(4G)}$$

Exercise 7.5

Transitivity of $<$ not only gives us the **4**-schema for G, but also the **4**-schema for H. Explain why.

Another plausible condition is that no time is earlier than itself; so $<$ should be *irreflexive*, meaning that we never have $t < t$. We know that reflexivity corresponds to principle **T**, whose (forward-looking) temporal analogue would be $\mathsf{G}\,A \rightarrow A$. What corresponds to irreflexivity? The following observation reveals the answer: nothing.

Observation 7.1: A sentence is valid in the class of irreflexive frames iff it is valid in the class of all frames.

Proof sketch: The right-to-left direction is obvious. For the left-to-right direction, suppose that some sentence A is not valid in the class of all frames. We need to show that A is not valid in the class of irreflexive frames. That A is not valid in the class of all Kripke frames means that there is some world w in some Kripke model $M = \langle W, R, V \rangle$ at which A is false. We have to show that there is some world in some irreflexive model at which A is false.

To this end, we will construct an irreflexive model $M^i = \langle W', R', V' \rangle$ from M in which the same sentences are true at w as in M. Since A is true at w in M, it follows that A is true at w in M^i.

Initially, M^i has the same worlds, the same accessibility relation, and the same interpretation function as M. Now for any world w in M that can see itself, we add a new world w' to M^i so that

- w' verifies the same sentence letters as w: for all ρ, $V'(\rho, w') = V'(\rho, w)$;
- w' can see the same worlds as w: whenever $wR'v$ then $w'R'v$; and
- w' can be seen from the same worlds as w: whenever $vR'w$ then $vR'w'$.

Finally, we make w inaccessible from itself in M^i. A simple proof by induction on complexity shows that if a sentence is true at a world w in M then it is also true at w in M^i. □

Observation 7.1 tells us that there is no modal principle that is valid in all and only the irreflexive frames. So the logic of irreflexive frames is the same as logic

of all frames (system K). The proof carries over to many other classes of frames. For example, the logic of irreflexive and transitive frames is the same as the logic of transitive frames (namely, K4).

Notice that we can easily add a tree rule for irreflexivity: simply allow any branch to be closed that contains $\omega < \omega$. But while that rule may help to find irreflexive countermodels, it usually won't allow us to prove anything we couldn't prove without the rule.

Given transitivity, irreflexivity is closely related to asymmetry. Recall from the previous chapter that $<$ is asymmetric if $t < s$ entails $s \not< t$. Asymmetry is also plausible if we interpret $<$ as the precedence relation between times. But as with irreflexivity, there is no modal schema that corresponds to asymmetry.

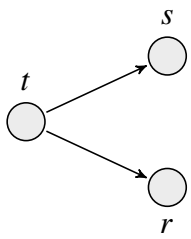
Exercise 7.6

Show that a transitive relation is irreflexive iff it is asymmetric.

Exercise 7.7

A popular idea in many cultures is that time is circular. Does this cast doubt on asymmetry? What about irreflexivity?

If $<$ is transitive and irreflexive (or equivalently: transitive and asymmetric), then it is a *partial order*. Partial orders are called "partial" because they don't necessarily order everything. For example, in a model of branching time we can have $t < s$ and $t < r$ but neither $s < r$ nor $r < s$; so r and s are not ordered by the precedence relation.



For many applications, we may want to rule out such cases, by imposing another requirement of **connectedness**. Connectedness demands that for any points $t, s \in T$, either $t < s$ or $t = s$ or $s < t$. (Note that this allows for cases where $t < s$ *and* $s < t$.) A relation that is irreflexive, transitive, and connected is called a **linear order**.

For other applications, we may want linearity in only one direction. Many philosophers have been attracted to a branching-future conception of time, on which a given point in time may have more than one future, but only one past. In such models, we would only require **left-linearity**: that *if* $s < t$ *and* $r < t$, then either

$s < r$ or $s = r$ or $r < s$.

The axiom schema corresponding to left-linearity is **BL** (for "backwards-looking linearity").

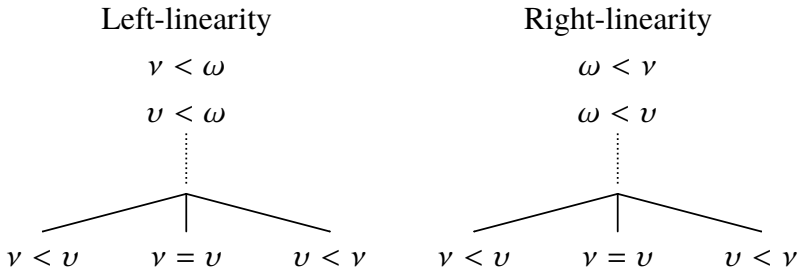
$$\mathsf{F}\,\mathsf{P}\,A \rightarrow (\mathsf{F}\,A \vee A \vee \mathsf{P}\,A) \qquad (\mathbf{BL})$$

Right-linearity similarly corresponds to **FL**.

$$\mathsf{P}\,\mathsf{F}\,A \rightarrow (\mathsf{P}\,A \vee A \vee \mathsf{F}\,A) \qquad (\mathbf{FL})$$

Right-linearity means that if $t < s$ and $t < r$, then either $s < r$ or $s = r$ or $r < s$. Together, left-linearity and right-linearity are equivalent to connectedness. So the conjunction of **BL** and **FL** corresponds to connectedness.

The tree rules for left-linearity and right-linearity are as you might expect from the definition of these two properties.



These rules create *three* branches. Also, they create nodes of the form $\nu = \upsilon$, stating that two world/time labels refer to the same thing. We need two further rules to deal with this kind of node. Both of these rules are called 'Identity'.

Exercise 7.8

Give tree proofs for the following schemas, assuming time is linear and transitive (i.e., using the Transitivity, Left-linearity, Right-linearity, and Identity rules).

(a) $\mathsf{P}\,\mathsf{G}\,A \rightarrow \mathsf{P}\,\mathsf{F}\,A$
(b) $\mathsf{P}\,\mathsf{G}\,\mathsf{G}\,A \rightarrow \mathsf{G}\,\mathsf{G}\,A$
(c) $\mathsf{P}\,\mathsf{F}\,A \rightarrow (\mathsf{P}\,A \vee (A \vee \mathsf{F}\,A))$
(d) $(\mathsf{F}\,A \wedge \mathsf{F}\,B) \rightarrow (\mathsf{F}(A \wedge B) \vee (\mathsf{F}(A \wedge \mathsf{F}\,B) \vee \mathsf{F}(\mathsf{F}\,A \wedge B)))$
(e) $\mathsf{F}(\mathsf{G}\,B \wedge \neg A) \rightarrow \mathsf{G}(A \rightarrow (\mathsf{G}\,A \rightarrow B))$

Exercise 7.9

Use the tree method to find countermodels for the following sentences, assuming time is linear and transitive.

(a) $(\mathsf{F}\,p \wedge \mathsf{F}\,q) \rightarrow \mathsf{F}(p \wedge q)$
(b) $\mathsf{F}\,p \rightarrow \mathsf{F}\,\mathsf{F}\,p$
(c) $\mathsf{P}\,\mathsf{H}\,p \rightarrow \mathsf{H}\,p$
(d) $\mathsf{F}\,\mathsf{G}\,A \rightarrow \mathsf{G}\,\mathsf{F}\,A$

The precedence relation in relativistic spacetime models is neither left-linear nor right-linear. But it has a weaker property that we already know: convergence. A spacetime point p_1 can precede two points p_2 and p_3 neither of which precedes the other, but these two points will always precede another point p_4. Convergence corresponds to the **G**-schema. In temporal logic, we have one **G**-schema for future convergence and one for past convergence.

$$\mathsf{F}\,\mathsf{G}\,A \rightarrow \mathsf{G}\,\mathsf{F}\,A \qquad \textbf{(FG)}$$
$$\mathsf{P}\,\mathsf{H}\,A \rightarrow \mathsf{H}\,\mathsf{P}\,A \qquad \textbf{(BG)}$$

Exercise 7.10

Can you find schemas that correspond to the following frame properties?

(a) There is no last time. (That is, every time precedes some time.)
(b) There is no first time.

(c) There is a last time.
(d) There is a first time.

Exercise 7.11

Show that the schema $\mathsf{F}\,A \to \mathsf{F}\,\mathsf{F}\,A$ corresponds to density.

Exercise 7.12

Can you define *Always* and *Sometimes* in terms of F, G, P, and H? Can you do so if you make assumptions about the flow of time $<$?

7.4 Branching time

It is natural to think of the future as "open" in a way that the past is "closed". I might decide to go for a walk this afternoon, or I might decide to stay at home. So there appear to be multiple futures: in some I go for a walk, in others I stay at home. One might conclude that time is left-linear, but not right-linear.

However, this argument is too quick. If I haven't decided what to do in the afternoon, then there are several *possible* futures – several ways the worlds *might* evolve. But it is far from clear that there are several *actual* futures. If there were, it would make little sense to wonder what I will do, or to contemplate whether I should go for a walk or stay at home: I would end up doing both anyway, albeit on different temporal branches.

A better way to capture the intuition that the future is open would use a multi-modal language with both temporal and circumstantial operators. The openness of the future could then be expressed by statements like $\Diamond\,\mathsf{G}\,p \wedge \Diamond\,\mathsf{G}\,\neg p$. To prevent a corresponding openness of the past, the accessibility relation for the circumstantial diamond would have to hold fixed the past, so that a world v is accessible from a world w only if the past of v coincides with the past of w.

On the other hand, there are also views that assume a genuinely branching flow of time. Some of these are motivated by discoveries in 20th century physics. I already mentioned that the precedence relation in relativistic spacetime allows for branching, although all these branches ultimately reconverge. A more classical form

of branching (without reconvergence) has been argued to follow from the so-called "Everett interpretation" of quantum physics. On this interpretation, what are normally understood to be chance events are really branching events in which all possible outcomes actually take place.

Earlier, I mentioned that branching time models are also widely used in computer science, where the "times" represent states of a computational process and the precedence relation holds between two states if the first can lead to the second. (The relevant logics of branching time are called "computational tree logics").

Another way to motivate a branching conception of time arises from a metaphysical view called **presentism**. According to presentism, only the present is real; all truths that seem to talk about other times are reducible to more fundamental truths about the present. For example, if it is true that there was a sea battle yesterday, then according to presentism this must ultimately be explained by what is true *now*; there must be facts about the present state of the world which entail that (and explain why) there was a sea battle yesterday. On one form of presentism, the relevant facts about the present state of the world are (a) particular facts about the distribution of physical particles and fields etc., and (b) the general laws of nature. Many laws of nature are dynamic, specifying how closed physical systems evolve over time. So the laws might entail that if the present physical state of the world is so-and-so, then there was a sea battle yesterday. This is how facts about the present might entail that there was a sea battle in the past.

But now suppose the laws of nature are indeterministic towards the future: they merely settle that if the present physical state of the world is so-and-so, then the future is *either like this or like that*. In that case, the presentist will deny that exactly one of these futures is actual.

Let's assume, then, that we want to reason about branching time. As we will see, this is less straightforward than it might at first appear.

The models we are interested in are not right-linear. I will, however, assume that they satisfy the following weaker property:

$$\text{if } t < s \text{ then for any } r, \text{ either } t < r \text{ or } r < s.$$

This is sometimes called "transitivity of nonprecedence", because it is equivalent to the assumption that if $t \not< s$ and $s \not< r$ then $t \not< r$. It slightly simplifies our models, for example by ruling out multiple parallel time lines which never connect to each

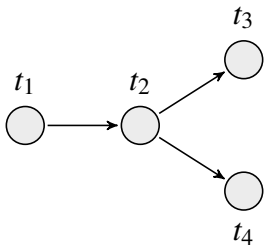
other.

Two pieces of terminology will be useful. First, let's define a **branch** in a model $\langle T, <, V \rangle$ as a maximal linearly ordered subset of T. (Branches are also known as "histories".) That is, a branch is a collection of times B such that

(a) for all t and s in B, either $t < s$ or $t = s$ or $s < t$, and

(b) no further member of T could be added to B without making (a) false.

The model (or rather, frame) depicted on the right, for example, contains two branches: $\{t_1, t_2, t_3\}$ and $\{t_1, t_2, t_4\}$.



Second, if t is any time in any model, then any maximal linearly ordered set of times *later than* t will be called a **future of** t. In the example on the right, t_1 has two futures: $\{t_2, t_3\}$ and $\{t_2, t_4\}$.

If you look back at definition 7.2, you can see that in the standard semantics for temporal logic, $\mathsf{G}\,p$ is true at t iff p is true at all times *in all futures of* t; $\mathsf{F}\,p$, on the other hand, is true at t iff p is true at some time *in at least one future of* t. This ensures that G and F are duals, but it may be regarded as problematic if we want $\mathsf{F}\,p$ to translate 'it will be the case that p'.

To illustrate, suppose I'm about to toss a coin. In one future, the coin will land heads, in another it will land tails. By definition 7.2, both $\mathsf{F}\,h$ and $\mathsf{F}\,t$ are true. But it is not clear if we should say that the coin will land heads and also that it will land tails.

One might therefore suggest an alternative semantics for F according to which $\mathsf{F}\,p$ is true at t iff p is true at some time in *all* futures of t:

$$M, t \models \mathsf{F}\,A \text{ iff every future of } t \text{ contains some } s \text{ such that } M, s \models A.$$

This is known as the **Peircean interpretation** of F (after Charles S. Peirce; the label is due to Arthur Prior).

On the Peircean account, $\mathsf{F}\,p$ is false if p only takes place in one of several futures. If we keep the classical interpretation of G, both $\mathsf{F}\,p$ and $\mathsf{G}\,\neg p$ can be false; so F and G are no longer duals. The dual of F is a strange operator that applies to a sentence A iff there is *some* future in which A is always true.

Exercise 7.13

Explain why Peirceanism renders **Con2** invalid.

A rather different approach is taken by (what Prior called) the **Ockhamist** approach. According to Ockhamism, if there are several futures, then it doesn't make sense to say – without qualification – that p will be the case, or that p won't be case. To talk about what will or won't be the case we must specify which future we have in mind.

Formally, in Ockhamist semantics, the truth-value of every sentence is evaluated at a pair consisting at a time and a branch. Branches are linear by definition, so the problems raised by multiple futures disappear. To say that p is the case in *some* branch, or in *all* branches, Ockhamists add new operators $\Diamond$ and $\Box$ which quantify over branches. So the Peircean F operator would be rendered $\Box\mathsf{F}$ in Ockhamism. $\Box\mathsf{F}\,p$ is true if p will eventually be true at some time in every future; $\Diamond Fp$, by contrast, would express that p is the case in some future.

Here is the full Ockhamist semantics.

Definition 7.3: Ockhamist Semantics

If $M = \langle T, <, V\rangle$ is a temporal model, B is a branch in M, t is a member of B, ρ is any sentence letter, and A, B are any sentences in the Ockhamist language, then

(a)	$M, B, t \models \rho$	iff $V(\rho, t) = 1$.
(b)	$M, B, t \models \neg A$	iff $M, B, t \not\models A$.
(c)	$M, B, t \models A \wedge B$	iff $M, B, t \models A$ and $M, B, t \models B$.
(d)	$M, B, t \models A \vee B$	iff $M, B, t \models A$ or $M, B, t \models B$.
(e)	$M, B, t \models A \rightarrow B$	iff $M, B, t \models B$ or $M, B, t \not\models A$.
(f)	$M, B, t \models A \leftrightarrow B$	iff $M, B, t \models (A \rightarrow B)$ and $M, B, t \models (B \rightarrow A)$.
(g)	$M, B, t \models \mathsf{F}\,A$	iff $M, B, s \models A$ for some $s \in B$ such that $t < s$.
(h)	$M, B, t \models \mathsf{G}\,A$	iff $M, B, s \models A$ for all $s \in B$ such that $t < s$.
(i)	$M, B, t \models \mathsf{P}\,A$	iff $M, B, s \models A$ for some $s \in B$ such that $s < t$.
(j)	$M, B, t \models \mathsf{H}\,A$	iff $M, B, s \models A$ for all $s \in B$ such that $s < t$.
(k)	$M, B, t \models \Box A$	iff $M, B', t \models A$ for all branches B' that contain t.
(l)	$M, B, t \models \Diamond A$	iff $M, B', t \models A$ for some branch B' that contains t.

A sentence is *valid* in Ockhamist semantics if it is true at all times on all branches in all models (ignoring branches that don't contain the relevant time). Nobody has yet found an axiomatic calculus that can prove all and only the Ockhamistically valid sentences.

Something else is strange about the Ockhamist approach. Informally, a sentence is logically true if it is true in every conceivable scenario under every interpretation of the non-logical vocabulary. Now consider a scenario in which there are multiple futures; one future holds a sea battle, another holds no sea battle. Let p translate 'there is a sea battle'. Is $\mathsf{F}\,p$ is true in this scenario (under the given interpretation of p)? What about $\mathsf{F}(p \vee \neg p)$? Or $Gp \rightarrow GGp$?

Ockhamism doesn't say. In Ockhamism, sentences are only true or false relative to a model and a time *and a branch*. Intuitively, however, a branching-time scenario does not include a particular branch. We'd like to know which sentences are true today if there are multiple futures. Ockhamism only tells us which sentences are true relative to each of the different futures: relative to a branch that contains a sea battle, $\mathsf{F}\,p$ is true, relative to other branches $\mathsf{F}\,p$ is false.

If we insist that logical validity should formalize the idea of truth in all scenarios under all interpretations of the non-logical vocabulary, then we can't accept the official definition of validity in Ockhamist semantics. We have to extend the Ockhamist semantics by specifying under what conditions a sentence is true *in a model at a time*, without fixing a branch. Then we can say that a sentence is valid iff it is true at all times in all models.

One simple strategy is to stipulate that a sentence is true at time in a model iff it is true relative to *all* branches that contain the time:

$$M, t \models A \text{ iff } M, B, t \models A \text{ for all branches } B \text{ that contain } t.$$

This is known as a **supervaluationist** semantics. A nice feature of supervaluationism is that a sentence comes out as true at all times in all models iff it is true at all times on all branches in all models. So the set of valid sentences is the same on the revised definition of validity (truth at all times in all models) as on the original definition (truth at all times on all branches in all models).

Many other answers are possible. For example, we could say that a sentence is true at a time in a model iff it is true relative to *some* branch containing the time. Or we could say that sentences can have three truth-values: a sentence is true if it is true

relative to all branches, false if it is false relative to all branches, and neither true nor false if it is true relative to some but not all branches. On this approach, 'there will be a sea battle' is neither true nor false if one future holds a sea battle and other futures do not.

Exercise 7.14

Which of the following schemas are valid in Ockhamist semantics, where 'valid' means true at all times on all branches in all models?

(a) $\Box A \rightarrow A$
(b) $\Box A \rightarrow \Box\Box A$
(c) $\Diamond A \rightarrow \Box\Diamond A$
(d) $\Box \mathsf{F} A \rightarrow \mathsf{F}\Box A$
(e) $\mathsf{P} A \rightarrow \Box \mathsf{P}\Diamond A$

7.5 Extending the language

The expressive resources of standard modal logic are weak. There are many things we might want to say about the unfolding of events in time that can't be said with F, G, P, and H. The Ockhamist branch quantifiers are one way of adding expressive power to the basic language of temporal logic. In this section, we will look at some others.

A useful operator for logics of discrete and linear time is the "next" operator X (also written '$\bigcirc$'). Informally, $\mathsf{X} A$ means that A is true at the next point in time. Formally:

$M, t \models \mathsf{X} A$ iff $M, s \models A$ for some s such that (a) $t < s$ and (b) $s < r$ for all r such that $t \neq s$ and $t < r$.

With the help of X, we can also say that A is true in two units of time: $\mathsf{X}\mathsf{X} A$, or that A is true in three units of time: $\mathsf{X}\mathsf{X}\mathsf{X} A$, and so on. The corresponding operator for talking about the *previous* point in time is usually written Y.

A much more powerful extension of $\mathfrak{L}_T$ adds binary operators for "since" and "until", which can be used to translate sentences like (1) and (2).

(1) Ever since we left the house it has been raining.

(2) It will be raining until we go back inside.

Informally, $\mathsf{S}(A, B)$ is true iff A was true at some time in the past and B has always been true since then; $\mathsf{U}(A, B)$ is true iff A will be true at some time in the future and B will always be true until then. Formally:

$M, t \models \mathsf{S}(A, B)$ iff there is some s with $s < t$ for which $M, s \models A$, and for all r with $s < r < t$, we have $M, r \models B$.

$M, t \models \mathsf{U}(A, B)$ iff there is some s with $t < s$ for which $M, s \models A$, and for all r with $t < r < s$, we have $M, r \models B$.

If we have S and U, we don't actually need $\mathsf{F}, \mathsf{G}, \mathsf{P}$, and H, because these can be defined in terms of S and U: $\mathsf{P}\, A$ is equivalent to $\mathsf{S}(A, p \vee \neg p)$, and $\mathsf{F}\, A$ to $U(A, p \vee \neg p)$. Conversely, however, $\mathsf{S}(A, B)$ and $\mathsf{U}(A, B)$ are not expressible in $\mathfrak{L}_t$.

Exercise 7.15

Define $\mathsf{X}\, A$ in terms of U.

Another noteworthy addition to temporal logic is the *Now* operator N. To see the point of this operator, consider the following multi-modal statement.

(3) Bob already knew yesterday that there would be a test today.

Using Y for 'yesterday', we might try to translate (3) as $\mathsf{Y}\, \mathsf{K}_b\, p$, where p translates 'there is a test'. But that's wrong. By the semantics for Y, $\mathsf{Y}\, \mathsf{K}_b\, p$ is true today iff $\mathsf{K}_b\, p$ is true yesterday (using days as temporal units). And since knowledge is factive, if $\mathsf{K}_b\, p$ is true at some time, then p is true at that time. So $\mathsf{Y}\, \mathsf{K}_b\, p$ wrongly entails that there was a test *yesterday*.

Intuitively, the problem is that 'today' in (3) refers to the present day, even though it occurs in the scope of 'yesterday'. The same thing happens in the quantified statement (4).

(4) One day everyone who is now rich will be poor.

Here, 'now' refers to the present time, even though it is in the scope of the F operator 'one day'.

With the *Now* operator N, (3) can be translated as $\mathsf{Y}\,\mathsf{K}_b\,\mathsf{N}\,p$, and (4) as $\mathsf{F}\,\forall x(\mathsf{N}\,Rx \rightarrow Px)$. (We will have a closer look at quantified modal logic in later chapters.)

The semantics of N raises a problem. Consider a simpler example, $\mathsf{P}\,\mathsf{N}\,p$. Since 'now' always picks out the present time, even when embedded under other temporal operators, $\mathsf{P}\,\mathsf{N}\,p$ should be equivalent to p (provided that the present time is not the first point in time). By the semantics of P,

$$M, t \models \mathsf{P}\,\mathsf{N}\,p \text{ iff } M, s \models \mathsf{N}\,p \text{ for some time } s < t.$$

Now we want $M, s \models \mathsf{N}\,p$ to be true iff p is true *at the original time* t. So we need to keep track of the original time at which we evaluate a sentence, even if a temporal operator shifts the time at which a subsentence is evaluated.

The standard way to achieve this is to define truth relative to *pairs* of times. One of the times is shifted by the modal operators, the other is held fixed.

Definition 7.4: Two-Dimensional Temporal Semantics

If $M = \langle T, <, V \rangle$ is a temporal model, t, t_0 are members of T, ρ is any sentence letter, and A, B are any $\mathfrak{L}_T$-sentences, then

(a)	$M, t_0, t \models \rho$	iff $V(\rho, t) = 1$.
(b)	$M, t_0, t \models \neg A$	iff $M, t_0, t \not\models A$.
(c)	$M, t_0, t \models A \wedge B$	iff $M, t_0, t \models A$ and $M, t_0, t \models B$.
(d)	$M, t_0, t \models A \vee B$	iff $M, t_0, t \models A$ or $M, t_0, t \models B$.
(e)	$M, t_0, t \models A \rightarrow B$	iff $M, t_0, t \models B$ or $M, t_0, t \not\models A$.
(f)	$M, t_0, t \models A \leftrightarrow B$	iff $M, t_0, t \models (A \rightarrow B)$ and $M, t_0, t \models (B \rightarrow A)$.
(g)	$M, t_0, t \models \mathsf{F}\,A$	iff $M, t_0, s \models A$ for some $s \in T$ such that $t < s$.
(h)	$M, t_0, t \models \mathsf{G}\,A$	iff $M, t_0, s \models A$ for all $s \in T$ such that $t < s$.
(i)	$M, t_0, t \models \mathsf{P}\,A$	iff $M, t_0, s \models A$ for some $s \in T$ such that $s < t$.
(j)	$M, t_0, t \models \mathsf{H}\,A$	iff $M, t_0, s \models A$ for all $s \in T$ such that $s < t$.
(k)	$M, t_0, t \models \mathsf{N}\,A$	iff $M, t_0, t_0 \models A$.

Like in Ockhamism, we also need to specify under what conditions a sentence is true (in a model) *at a time*, not at a pair of two times. Here, the standard proposal is

not supervaluationist but "diagonalist":

$$M, t \models A \text{ iff } M, t, t \models A.$$

Now we can show that $\mathsf{P}\,\mathsf{N}\,p$ is equivalent to p, provided there are earlier times. I will go through the left-to-right direction.

1. Assume $M, t \models \mathsf{P}\,\mathsf{N}\,p$.
2. Then $M, t, t \models \mathsf{P}\,\mathsf{N}\,p$, by the definition of truth at a time in a model.
3. Then $M, t, s \models \mathsf{N}\,p$ for some $s < t$, by clause (i) of definition 7.4.
4. Then $M, t, t \models p$ for some $s < t$, by clause (k) of definition 7.4.
5. Then $M, t \models p$ for some $s < t$, by the definition of truth at a time in a model.

The presence of a *Now* operator has far-reaching consequences for the logic of time. Note, for example, that $\mathsf{N}\,p \rightarrow p$ is valid, in the sense that it is true at all times in all models. By contrast, $\mathsf{G}(\mathsf{N}\,p \rightarrow p)$ is invalid: if p is true at t and false at some time after t, then $\mathsf{G}(\mathsf{N}\,p \rightarrow p)$ is false at t. So we must give up the forward and backward Necessitation rules. The fact that something is logically valid no longer entails that it will always be the case.

Exercise 7.16

Suppose we add to the language of basic temporal logic with N another operator $\Box$ that applies to a sentence iff the sentence is valid. Explain why at least one of the following assumptions must then be false. Which of them do you think we should reject?

(a) Whenever $\Box A$, then $\mathsf{G}\,\Box A$.
(b) For any A, $\mathsf{G}(\Box A \rightarrow A)$.

8 Conditionals

8.1 Material conditionals

A good thing about sentences of the form $A \to B$ (called **material conditionals**) is that their meaning is completely clear. $A \to B$ is true iff A is false or B is true. It is not so clear whether 'if A then B' sentences in natural language can be adequately translated as $A \to B$.

The following facts about logical consequence (in classical propositional logic) are widely thought to show that English conditionals are not material conditionals.

$$B \models A \to B \tag{\textbf{P1}}$$

$$\neg A \models A \to B \tag{\textbf{P2}}$$

$$\neg(A \to B) \models A \tag{\textbf{P3}}$$

$$A \to B \models \neg B \to \neg A \tag{\textbf{P4}}$$

$$A \to B \models (A \wedge C) \to B \tag{\textbf{P5}}$$

(P1)–(P5) are sometimes called "paradoxes of material implication", because they sound wrong if we assume that $A \to B$ translates 'if A then B' or (worse) 'A implies B'. Here are apparent counterexamples for each schema.

1. The lecture ends at 3pm. Therefore: If the building collapses at 2.45 then the lecture ends at 3pm.
2. The President won't be impeached. Therefore: If the President will be impeached then nobody will care.
3. It is not the case that if it will rain tomorrow then the Moon will fall onto the Earth. Therefore: It will rain tomorrow.
4. If our opponents are cheating, we will never find out. Therefore: If we will find out that our opponents are cheating, then they aren't cheating.

5. If you add sugar to your coffee, it will taste good. Therefore: If you add sugar and vinegar to your coffee, it will taste good.

Not everyone is convinced by examples like these. Some have argued that the conclusions really do follow from the premises, and have appealed to pragmatic explanations of why the inferences seem wrong.

There are also direct arguments in favour of the interpretation of English conditionals as material conditionals. For example, suppose I make the following promise.

(1) If I don't have to work tomorrow, then I will help you move.

Under what conditions will I have broken my promise? Clearly, I have made a false promise if the next day I don't have to work and yet I don't help you move. Under all other conditions, however, you couldn't fault me for having broken my promise. So it seems that (1) is false iff I don't have to work and I don't help you move. Generalizing, this suggests that 'if A then B' is false iff A is true and B is false. And then 'if A then B' is equivalent to $A \to B$.

Another argument starts with the intuitively plausible assumption that 'A or B' entails the corresponding conditional 'if not-A then B'. (This is sometimes called the *or-to-if principle*.) For instance, if I tell you that Nadia is either in Rome or in Paris, you can infer that if she's not in Rome then she's in Paris. Now we can reason as follows.

1. Suppose 'not-A or B' is false.
2. Then A is true and B is false.
3. Then the conditional 'if A then B' is clearly false.
4. So the falsehood of 'not-A or B' entails the falsehood of 'if A then B'.
5. By the or-to-if principle, the truth of 'not-A or B' entails the truth of 'if A then B'.
6. So 'if A then B' is logically equivalent to 'not-A or B', which is equivalent to $A \to B$.

Many more arguments have been given for and against the hypothesis that natural-language conditionals are material conditionals. We won't look further into this

debate. Even those who defend the reading of English conditionals as material conditionals admit that it does not work for all if-then sentences in natural language.

One kind of counterexample are generic conditionals like (2).

(2) If water is heated to 100° C, it evaporates.

This is plausibly a quantified (or modal) statement. The claim is that *in all (normal) cases* where water is heated to 100° C, it evaporates.

Another kind of counterexample are so-called subjunctive conditionals. Compare the following two statements.

(3) If Shakespeare didn't write *Hamlet*, then someone else did.

(4) If Shakespeare hadn't written *Hamlet*, then someone else would have.

(3) seems true. Someone has written *Hamlet*; so if it wasn't Shakespeare then it must have been someone else. But (4) is almost certainly false. After all, it is very likely that Shakespeare did write *Hamlet*. And it is highly unlikely that if he hadn't written *Hamlet* – if he got distracted by other projects, say – then someone else would have stepped in to write the exact same piece.

Sentences like (3) are called **indicative conditionals**; (4) is a **subjunctive conditional**. (Subjunctive conditionals are also called 'counterfactual conditionals' or 'counterfactuals'.)

Whatever we say about indicative conditionals like (3), subjunctive conditionals clearly can't be translated as material conditionals. As I just said, on the assumption that Shakespeare wrote *Hamlet*, (4) is almost certainly false, even though it has a false antecedent, and so the corresponding material conditional is true.

To sum up, there seem to be different kinds of conditionals – different kinds of if-then sentences – in natural language. At least some of them cannot be translated as material conditionals. If we want to formalize reasoning with these conditionals, we need a better translation.

8.2 Strict conditionals

A material conditional $A \to B$ can be true even though there is no connection between the antecedent A and the consequent B. If-then sentences in natural language seem different. Consider (1).

(1) If we leave after 5, we will miss the train.

Intuitively, if someone utters (1), they want to convey that missing the train is a *necessary consequence* of leaving after 5 – that it is impossible to leave after 5 and still make it to the train (given certain facts about the distance to the station, the time it takes to get there, etc.). This suggests that (1) should be formalized not as $p \rightarrow q$ but as $\Box(p \rightarrow q)$ or, equivalently, $\neg\Diamond(p \wedge \neg q)$.

Sentences of the form $\Box(A \rightarrow B)$ or $\neg\Diamond(A \wedge \neg B)$ are called **strict conditionals**. The label goes back to C.I. Lewis (1918), who also introduced the abbreviation $A \dashv\!\!\models B$ for strict conditionals.

Lewis was not interested in the interpretation of ordinary-language conditionals. He wanted $A \dashv\!\!\models B$ to formalize '*A* implies *B*' or '*A* entails *B*'. His intended use of $\dashv\!\!\models$ therefore roughly matches our use of the double-barred turnstile '$\models$'. But there are important differences. The turnstile is an operator in the *meta-language* we use to talk about sentences in $\mathfrak{L}_M$ and other formal languages. Lewis's $\dashv\!\!\models$, by contrast, is an *object-language* operator like $\wedge$ or $\rightarrow$ that can be placed between any two sentences in a formal language to generate another sentence in the language. For example, $p \dashv\!\!\models (q \dashv\!\!\models p)$ is well-formed, whereas $p \models (q \models p)$ is gibberish. Moreover, while $p \models q$ is simply false – because there are models in which p is true and q false – Lewis's $p \dashv\!\!\models q$ is true on some interpretation of the sentence letters and false on others. For instance, if p means that it is snowing and q that precipitation occurs, then $p \dashv\!\!\models q$ is plausibly true, because snowfall is a form of precipitation, so the hypothesis that it is snowing implies that precipitation occurs.

Let's set aside Lewis's project of formalizing the intuitive concept of implication. Our goal is to define an object-language operator that functions like 'if . . . then . . . ' in English. To see whether Lewis's $\dashv\!\!\models$ can do the job, we need to have a closer look at what it means.

Since $A \dashv\!\!\models B$ is equivalent to $\Box(A \rightarrow B)$, standard Kripke semantics for the box also provides a semantics for strict conditionals. In Kripke semantics, $\Box(A \rightarrow B)$ is true at a world w iff $A \rightarrow B$ is true at all worlds v accessible from w. And $A \rightarrow B$ is true at v iff either A is false at v or B is true at v. So we get the following truth-conditions for strict conditionals.

Definition 8.1: Kripke semantics for $\dashv\!\exists$

If $M = \langle W, R, V \rangle$ is a Kripke model, then
$M, w \models A \dashv\!\exists B$ iff for all v such that wRv, either $M, v \not\models A$ or $M, v \models B$.

Exercise 8.1

We can define $\dashv\!\exists$ in terms of $\Box$ and $\rightarrow$. Can you define $\Box$ in terms of $\dashv\!\exists$ and truth-functional operators? That is, can you find a sentence schema with $\dashv\!\exists$ as the only non-truth-functional operator that is equivalent (in Kripke semantics) to $\Box A$?

As always, the logic of strict conditionals depends on what constraints we impose on the accessibility relation. For example, without any constraints, $\dashv\!\exists$ does not validate *modus ponens*, in the sense that $A \dashv\!\exists B$ and A together do not entail B. We can easily see this by translating $A \dashv\!\exists B$ back into $\Box(A \rightarrow B)$ and setting up a tree.

1. $\Box(A \rightarrow B)$ (w) (Ass.)
2. A (w) (Ass.)
3. $\neg B$ (w) (Ass.)

With the K-rules, where we don't make any assumptions about the accessibility relation, there is nothing we can do with this tree.

Exercise 8.2

Give a countermodel in which $p \dashv\!\exists q$ and p are true at some world while q is false.

On the other hand, if we assume that the accessibility relation is reflexive, the tree closes:

4. wRw (Ref.)
5. $A \rightarrow B$ (w) (1,4)

6. $\neg A$ (w) (5) — x
7. B (w) (5) — x

It is not hard to show that *modus ponens* for $\multimap$ is valid on all and only the reflexive frames. So reflexivity is precisely what we need to render *modus ponens* valid.

Exercise 8.3

Confirm the following claims, by translating $A \multimap B$ into $\Box(A \to B)$.

(a) $\models_K A \multimap A$
(b) $\neg B \multimap \neg A \models_K A \multimap B$
(c) $A \multimap B, B \multimap C \models_K A \multimap C$
(d) $(A \vee B) \multimap C \models (A \multimap C) \wedge (B \multimap C)$
(e) $A \multimap (B \multimap C) \models_T (A \wedge B) \multimap C$
(f) $A \multimap B \models_{S4} C \multimap (A \multimap B)$
(g) $((A \multimap B) \multimap C) \multimap (A \multimap B) \models_{S5} A \multimap B$

If we want $A \multimap B$ to translate 'if A then B', we probably want *modus ponens* to be valid. So we'll want the relevant Kripke models to be reflexive. Along the same lines, we could now look at other conditions on the accessibility relation and decide whether they should be imposed based on what they imply for the logic of conditionals. But let's take a shortcut.

Above I suggested that sentence (1) can be understood to say that it is *impossible* to leave after 5 and still make it to the train. Impossible in what sense? There are many possible worlds at which we leave after 5 and still make it to the train; for example, worlds at which the train departs two hours later, or worlds at which we live right next to the station. When I say that it is impossible to leave after 5 and still make it to the train, I mean that it is impossible *given what we know about the departure time, our location, etc.*

Generalizing, a tempting proposal is that for indicative conditionals like (1), the accessibility relation is the epistemic accessibility relation we studied in chapter 5: a world is accessible from w iff it is compatible with what is known at w. The logic of indicative conditionals is then determined by the logic of epistemic necessity; we don't need to figure out the relevant accessibility relations from scratch.

Since knowledge varies from agent to agent, the present idea implies that the truth-value of indicative conditionals should be agent-relative. This seems to be confirmed by the following puzzle, due to Allan Gibbard (1981).

> Sly Pete and Mr. Stone are playing poker on a Mississipi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point the room is cleared. A few minutes later, Zack slips me a note which says 'if Pete called, he won', and Jack slips me a note which says 'if Pete called, he lost'.

The puzzle is that Zack's note and Jack's note are intuitively contradictory, yet they both seem to be true.

We can resolve the puzzle if we understand the conditionals as strict conditionals with an agent-relative epistemic accessibility relation. Take Zack. Zack knows that Pete knows Stone's hand. Knowing that Pete would not call unless his hand is better, among the worlds compatible with Zack's knowledge, all worlds at which Pete calls are worlds at which Pete wins. So if p translates 'Pete called' and q 'Pete won', then $p \dashrightarrow q$ is true relative to Zack's information state. Relative to Jack's information state, however, the same sentence is false. Jack knows that Stone's hand is better than Pete's, but he doesn't know that Pete knows Stone's hand. So among the worlds compatible with Jack's knowledge, all worlds at which Pete calls are worlds at which Pete loses. Relative to Jack's information state, $p \dashrightarrow \neg q$ is true.

Another advantage of the "epistemically strict" interpretation of indicative conditionals is that it explains why indicative conditionals with antecedents that are known to be false seem defective. For example, suppose Fred has gone to work. In that scenario, is (2) true or false?

(2) If Fred has not gone to work, he is helping his neighbours.

The question is hard to answer, and not because we lack information about the scenario. Once we are told that Fred has gone to work, it is unclear how we are meant to assess whether Fred is helping his neighbours *if* he has not gone to work. On the epistemically strict interpretation, if A is known to be false, no A-world is epistemically accessible, and so it is pointless to ask whether all accessible A-worlds are B-worlds.

So there are some reasons to think that indicative conditionals express strict conditionals with an epistemic accessibility relation. What about subjunctive conditionals? Return to the two Shakespeare conditionals from the previous section.

When we evaluate the indicative sentence – 'If Shakespeare didn't write *Hamlet*, then someone else did' – we hold fixed our knowledge that *Hamlet* exists; worlds where the play was never written are inaccessible. But when we evaluate the subjunctive sentence – 'If Shakespeare hadn't written *Hamlet*, then someone else would have' – we don't hold fixed the fact that *Hamlet* exists. Otherwise the conditional would be true. So the accessibility relation for subjunctive conditionals does not track our knowledge or information. Unfortunately, as we are going to see in the next section, it is hard to say what else it tracks, because that seems to vary from conditional to conditional.

This is one problem for the strict analysis of natural-language conditionals. Another problem lies in the logic of strict conditionals. Let's have another look at the "paradoxes of material implication" from page 147. The strict analogues of **P1**–**P5** would go as follows:

$$B \models A \dashv\!\!\!\supset B \qquad \textbf{(P1S)}$$

$$\neg A \models A \dashv\!\!\!\supset B \qquad \textbf{(P2S)}$$

$$\neg(A \dashv\!\!\!\supset B) \models A \qquad \textbf{(P3S)}$$

$$A \dashv\!\!\!\supset B \models \neg B \dashv\!\!\!\supset \neg A \qquad \textbf{(P4S)}$$

$$A \dashv\!\!\!\supset B \models (A \wedge C) \dashv\!\!\!\supset B \qquad \textbf{(P5S)}$$

Of these, **P1S**–**P3S** are easily seen to be false (unless we require that every world has only access to itself, which would make $\Box(A \to B)$ equivalent to $A \to B$). But **P4S** and **P5S** are true, no matter what we say about accessibility.

So if we want to faithfully formalize ordinary-language conditionals, we either have to explain away the apparent counterexamples to 4 and 5, or find another translation.

Exercise 8.4

Show that **PS4** and **PS5** are true on the Kripke semantics for $\dashv\!\!\!\supset$ (for example, by giving a tree proof).

Exercise 8.5

In section 1, I gave examples showing that **P4** and **P5** sound wrong if indicative conditionals are translated as material conditionals (and so **P4S** and **P5S** sound wrong if indicative conditionals are translated as strict conditionals). Show that the problem also arises for subjunctive conditionals.

Exercise 8.6

A plausible norm of pragmatics is that a sentence should only be asserted if it is known to be true. If the logic of knowledge is at least S4, it follows that an epistemically strict conditional is assertable iff the corresponding material conditional is assertable. Explain.

8.3 Variably strict conditionals

I mentioned that the strict interpretation of conditionals has a problem with subjunctive conditionals. (In fact, the problem also arises for indicative conditionals, but it is easier to see with subjunctives.) The problem is best explained by an example.

As I am writing these notes, I am in Coombs Building, room 2228, with my desk facing the wall to Al Hájek's office in room 2229. In that context, (1) seems true.

(1) If I were to drill a hole through the wall behind my desk, the hole would come out in Al's office.

Once again, there is no logical connection between the antecedent of (1) and the consequent. There are many possible worlds at which I drill a whole through the wall behind my desk and don't come out in Al's office – for example, worlds at which my desk faces the opposite wall, worlds at which Al's office is in a different room, and so on. If we want to translate (1) as a strict conditional, all such worlds must be inaccessible.

Now consider (2).

(2) If the office spaces had been randomly reassigned yesterday, Al's office would (still) be next to mine.

(2) seems false, or at least extremely unlikely. But if we hold fixed that I am in room 2228 and that Al is in 2229 – as we seem to do for (1) – then (2) should be true.

Among the worlds at which I am in 2228 and Al is in 2229, all worlds at which the office spaces have been randomly reassigned yesterday are worlds where Al's office is next to mine. So when we evaluate (2), we no longer hold fixed who is in which office. Worlds that were inaccessible for (1) are accessible for (2).

So the accessibility relation for subjunctive conditionals appears to vary from conditional to conditional. As David Lewis put it, subjunctive conditionals are not strict, but "variably strict".

Let's try to get a better grip on how this might work. (What follows is a slightly simplified version of an analysis developed by Robert Stalnaker and David Lewis at around 1970.)

Intuitively, when we ask what would have been the case if a certain event had occurred, we are looking at worlds that are much like the actual world up to the time of the event, then deviate in some minimal way to allow the event to take place, and from then on unfold in accordance with the general laws of the actual world. For example, when we ask what would have happened if Shakespeare hadn't written *Hamlet*, we wonder what happens at worlds that are much like the actual world until 1599, at which point some mundane circumstance prevented Shakespeare from writing *Hamlet*. Similarly for (1). Here we are looking at worlds that are much like the actual world up to now, at which point I decide to drill a hole and find a suitable drill. These changes do not require my office to be in a different room, so worlds where I'm not in room 2228 are ignored. Figuratively speaking, such worlds are "too remote": they differ from the actual world in ways that are not required to make the antecedent true.

On that picture, a subjunctive conditional is true iff the consequent is true at the *closest* worlds at which the antecedent is true – where closeness is a matter of similarity in certain respects. There is a large literature on how the relevant similarity standards might be spelled out. Let's ignore this issue, as it is largely irrelevant to logic. Instead, let's simply assume that we have a suitable similarity ordering $\prec$ between possible worlds. So '$v \prec_w u$' means that v is closer to w than u, in the sense that v differs less than u from w in whatever respects are relevant to the interpretation of subjunctive conditionals.

We make the following structural assumptions about $\prec$.

1. If $v \prec_w u$ then $u \not\prec_w v$. (Asymmetry)
2. If $v \prec_w u$ and $u \prec_w t$ then $v \prec_w t$. (Transitivity)

3. If $v \prec_w u$, then for all t either $v \prec_w t$ or $t \prec_w u$. (Transitivity of $\not\prec_w$)
4. For all worlds v and u, $v \not\prec_w w$. (**Weak centring**)
5. For any non-empty set of worlds X and world w there is a world $v \in X$ such that there is no world $u \in X$ with $u \prec_w v$. (The **Limit Assumption**)

The first two are self-explanatory. The third assumption ensures that the "equidistance" relation that holds between u and v if neither $u \prec_w v$ nor $v \prec_w u$ is an equivalence relation. With these three assumptions, we can picture each world w as surrounded by nested spheres of other worlds; $u \prec_w v$ means that u is in a closer sphere around w than v.

Weak centring (assumption 4) means that every world is among the closest worlds to itself. Finally, the Limit Assumption ensures that for any consistent proposition A and world w, there is a set of closest A-worlds. Without the Limit Assumption, there could be an infinite chain of ever closer A-worlds, with no world being maximally close.

Exercise 8.7

Define a weaker relation $\preceq$ so that $u \preceq_w v$ iff $v \not\prec_w u$. Informally, $u \preceq_w v$ means that v is at least as similar to w in the relevant respects as u. Can you express the above five conditions on $\prec$ in terms of $\preceq$?

We now introduce a variably strict operator $\boxrightarrow$ so that $A \boxrightarrow B$ is true iff B is true at the closest worlds at which A is true. Any model for a language that contains the $\boxrightarrow$ operator must contain a closeness ordering $\prec$ on the set of worlds.

Definition 8.2

A **similarity model** consists of

- a non-empty set W,
- for each $w \in W$ an order $\prec_w$ that satisfies the above five conditions, and
- a function V that assigns to each sentence letter and each member of W a truth-value.

To formally specify the semantics of $A \boxrightarrow B$, define $Min_M^{\prec_w}(A)$ to be the set of closest A-worlds to w in model M. That is, for any sentence A and any world w in

any model M,

$$Min_M^{\prec_w}(A) =_{\text{def}} \{v : M, v \models A \text{ and } \neg\exists u(u \prec_w v \text{ and } M, w \models A)\}.$$

Then we can give the following truth-conditions for $A \boxrightarrow B$.

Definition 8.3: Similarity semantics for $\boxrightarrow$

If M is a similarity model and w a world in M, then
$M, w \models A \boxrightarrow B$ iff $M, v \models B$ for all $v \in Min_M^{\prec_w}(A)$.

To see this in action, let's verify that *modus ponens* is valid for $\boxrightarrow$. That is, we want to show that if $A \boxrightarrow B$ and A are both true at some world w in some similarity model, then so is B. If $A \boxrightarrow B$ is true at w, then by definition 8.3, B is true at all worlds in $Min_M^{\prec_w}(A)$. By the fact that A is true at w and weak centring, w is in $Min_M^{\prec_w}(A)$. So B is true at w.

Exercise 8.8

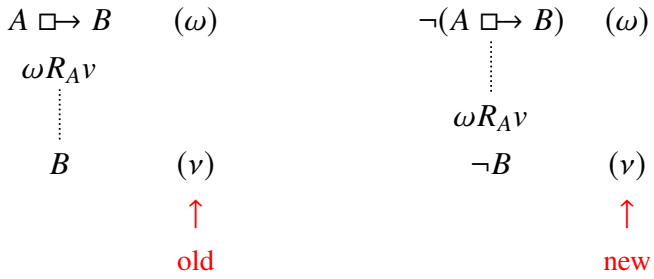
Explain why $A \boxrightarrow B$ entails $A \rightarrow B$.

Exercise 8.9

Show that if A is true at no worlds, then $A \boxrightarrow B$ is true.

You may have noticed that definition 8.3 resembles the classical Kripke semantics for $\Box B$. To make the resemblance more explicit, let's say that for any sentence A, a world v is *A-accessible* from w (for short, wR_Av) iff v is one of the closest A-worlds to w; that is, iff $v \in Min_M^{\prec_w}(A)$. Then definition 8.3 states that $A \boxrightarrow B$ is true at a world w iff B is true at all worlds A-accessible from w. These are the standard truth-conditions for the box, except that accessibility is relativised to the antecedent A.

These observations show that we can adapt the standard tree rules for the box to reason with $\boxrightarrow$, as follows.



To get a complete tree system, we need further rules. For example, the above two rules don't account for the fact that the closest A-worlds are always A-worlds. That is, if a world v is R_A-accessible from w, then A is true at v. They also don't account for the fact that every A-world is R_A-accessible to itself (by weak centring). Let's add two more rules to fill these gaps.

These rules are still not complete, because they don't reflect various interactions between the different accessibility relations. For example, if the worlds that are A-accessible from some world w include B-worlds, then definition 8.3 implies that the $A \wedge B$-accessible worlds from w are contained within the worlds A-accessible from w. The complete tree rules for the $\boxrightarrow$ operator are rather complicated. I will leave it at the above four rules, which suffice to establish many useful facts about variably strict conditionals.

Here, for illustration, is a tree proof to show once more that *modus ponens* is valid.

1.	$A \boxrightarrow B$	(w)	(Ass.)
2.	A	(w)	(Ass.)
3.	$\neg B$	(w)	(Ass.)
4.	$wR_A w$		(Centring)
5.	B	(w)	(1,4)
	x		

Exercise 8.10

Give tree proofs for the following statements.

(a) $A, \neg B \models \neg(A \boxright B)$
(b) $A \boxright B, A \boxright C \models A \boxright (B \wedge C)$
(c) $A \boxright (B \wedge C) \models (A \boxright B) \wedge (A \boxright C)$
(d) $A \boxright \neg A \models A \boxright B$

Since the tree rules I have presented are sound, you can be sure that whenever a tree closes then the tested entailment or validity holds. But since the rules are not complete, care is required when a tree doesn't close. You always need to check if a countermodel constructed from an open tree is an actual countermodel.

Constructing countermodels from open trees is at any rate not entirely straightforward. By way of illustration, let's show that $A \boxright B$ and $B \boxright C$ does not entail $A \boxright C$. The tree starts like this.

1.	$A \boxright B$	(w)	(Ass.)
2.	$B \boxright C$	(w)	(Ass.)
3.	$\neg(A \boxright C)$	(w)	(Ass.) ✓
4.	wR_Av		(3)
5.	$\neg C$	(v)	(3)
6.	B	(v)	(1,4)
7.	A	(v)	(4,Truth)

The Centring rule would allow us to add six more lines, but they wouldn't be useful, so let's stop here.

The open tree suggests that there is a countermodel with two worlds, w and v. At v, A and B are true. We also have wR_Av, so v is among the closest A-worlds to w. Since wR_Aw is not in the tree, we assume that w is not among the closest A-worlds to w, which also means that A is false at w. We also don't have wR_Bv. So even though B is true at v, v is not among the closest B-worlds to w. We can ensure this by assuming that B is true at w itself, and that w is the unique closest B-world from itself. Now you can verify that $A \boxright B$ and $B \boxright C$ are both true at w while $A \boxright C$ is false.

The example also illustrates one of the many differences between $\boxright$ and $\thickspace\!\!\!\!\sim$: while $A \thickspace\!\!\!\!\sim B$ and $B \thickspace\!\!\!\!\sim C$ entail $A \thickspace\!\!\!\!\sim C$, the same inference with $\boxright$ is invalid. And

arguably the inference *is* invalid with subjunctive conditionals. Stalnaker gives the following (cold-war era) counterexample.

> If Hoover had been born a Russian, he would have been a Communist.
> If Hoover were a Communist, he would have been be a traitor.
> Therefore, If Hoover had been born a Russian, he would have been be a traitor.

Exercise 8.11

Give counterexamples to the following, either by trying to construct them from open trees, or directly.

(a) $B \models A \boxrightarrow B$
(b) $\neg A \models A \boxrightarrow B$
(c) $\neg(A \boxrightarrow B) \models A$
(d) $A \boxrightarrow B \models \neg B \boxrightarrow \neg A$
(e) $A \boxrightarrow B \models (A \wedge C) \boxrightarrow B$

As the preceding exercise shows, none of the "paradoxes of material implication" carry over to variably strict conditionals. In this respect, $\boxrightarrow$ seems to better match the ordinary use of conditionals than $\supset$. On the other hand, you might have thought that the following entailment facts should hold, yet they do not. (The corresponding inferences with $\supset$ are valid; see exercise 8.2.)

1. $((A \vee B) \boxrightarrow C) \models (A \boxrightarrow C) \wedge (B \boxrightarrow C)$
2. $A \boxrightarrow (B \boxrightarrow C) \models (A \wedge B) \boxrightarrow C$

The semantics I have presented for $\boxrightarrow$ is a middle ground between those of Lewis and Stalnaker. Stalnaker assumes that $\prec_w$ satisfies the further assumption of linearity: either $w_1 \prec w_2$ or $w_1 = w_2$ or $w_2 \prec w_1$. Informally, Stalnaker thereby rules out ties in similarity. This renders the following principle valid, which is known as *Conditional Excluded Middle*.

$$(A \boxrightarrow B) \vee (A \boxrightarrow \neg B) \qquad \textbf{(CEM)}$$

There is an ongoing controversy over whether subjunctive conditionals in natural language validate conditional excluded middle. On the one hand, it is natural think

that 'it is not the case that if A then B' entails 'if A then not B'. On the other hand, suppose I have a number of coins in my pocket, none of which I have tossed. What would have happened if I had tossed one? Arguably, I might have gotten heads and I might have gotten tails. Either result is possible, but neither *would* have come about.

Exercise 8.12

Show that the following statements are true on Stalnaker's semantics:

(a) $A \wedge B \models A \boxrightarrow B$
(b) $A \boxrightarrow (B \vee C) \models (A \boxrightarrow B) \vee (A \boxrightarrow C)$

Lewis not only rejects linearity, but also the Limit Assumption, arguing that there might well be an infinite chain of ever closer A-worlds. Definition 8.3 implies that if there are no closest A-worlds then any sentence of the form $A \boxrightarrow B$ is true. That does not seem right. Lewis therefore gives a more complicated semantics:

> $M, w \models A \boxrightarrow B$ iff either there is no v for which $M, v \models A$ or there is some world v such that $M, v \models A$ and for all $u \prec_w v$, $M, w \models A \rightarrow B$.

It turns out that it makes no difference to the logic whether we impose the Limit Assumption and use the old definition or don't impose the Limit Assumption and use Lewis's new definition. The same sentences are valid either way.

8.4 The restrictor analysis

In section 6.3, we looked at conditional obligation statements like (1).

(1) If Jones is going to help his neighbours then he ought to tell them he's coming.

I claimed that this is best formalized not as $p \rightarrow \mathsf{O}\, q$ or $\mathsf{O}(p \rightarrow q)$, but as $\mathsf{O}(q/p)$, where $\mathsf{O}(\cdot/\cdot)$ is a primitive two-place operator for conditional obligation. However, the original sentence (1) appears to contain a conditional ('if . . . then . . . ') and the ordinary one-place operator 'ought'. One would like to know how these components work together to determine the intuitive meaning of (1).

To this end, we might reconsider if (1) can't be formalized with the help of a strict or variably strict conditional, say, as $\mathsf{O}(p \boxrightarrow q)$. Here I will explore the opposite

strategy, of analysing *all* conditionals on the model of conditional obligation. This turns out to work so well that it has become the standard approach in linguistics.

Let's begin with a rather different case, to which David Lewis drew attention in 1975. Consider (2) and (3).

(2) If it rains, we always stay inside.

(3) If it rains, we sometimes stay inside.

Let A mean 'always' and S 'sometimes'. How could we translate (2) and (3)?

The "narrow scope" translations $r \rightarrow \mathsf{A}\, s$ and $r \rightarrow \mathsf{S}\, s$ are easily seen to be inadequate. For (2), the "wide scope" $\mathsf{A}(r \rightarrow s)$ looks good. One might expect that (3) should then be translated as $\mathsf{S}(r \rightarrow s)$. But that is clearly wrong. (Note that $\mathsf{S}(r \rightarrow s)$ is equivalent to $\mathsf{S}(\neg r \vee s)$.) Intuitively, (3) means that there are times at which it rains *and* we stay inside. So it should be translated as $\mathsf{S}(r \wedge s)$. This is a bit surprising, because (3) seems to contain a conditional, yet in the translation we have a conjunction.

Things get worse if we turn to (4).

(4) If it rains, we mostly stay inside.

This says that among the occasions on which it rains, most are occasions on which we stay inside. Let M translate 'Mostly'. Neither $\mathsf{M}(r \rightarrow s)$ nor $\mathsf{M}(r \wedge s)$ capture the truth-conditions of (4), nor do $\mathsf{M}(r \mathrel{\Box\!\!\rightarrow} s)$ or $\mathsf{M}(r \mathrel{\neg\!\exists} s)$. Indeed, no statement of the form $\mathsf{M}\, A$ correctly translates (4).

We can formalize (4) if we treat M as a binary operator, taking two propositions as arguments: $\mathsf{M}(s/r)$. The semantics for the two forms of M would look as follows, assuming that we are quantifying over times.

$$
\begin{array}{ll}
M, t \models \mathsf{M}\, A & \text{iff } M, s \models A \text{ for most times } s. \\
M, t \models \mathsf{M}(A/B) & \text{iff } M, s \models A \text{ for most times } s \text{ such that } M, s \models B.
\end{array}
$$

So $\mathsf{M}\, A$ says that most times are A-times, while $\mathsf{M}(A/B)$ says that *among B-times*, most times are A-times. The function of the second argument to $\mathsf{M}(\cdot/\cdot)$, which translates the 'if'-clause in (4), is to **restrict the domain** of times over which the operator M quantifies.

If we assume that the 'if'-clause in (4) is a restrictor of 'mostly', it is natural to assume that it does the same thing in (2) and (3). (3) states that *among times when*

it rains, we sometimes stay inside, while (2) states that among the same times, we always stay inside.

Exercise 8.13

How would you translate 'all dogs are barking', 'some dogs are barking', and 'most dogs are barking' into predicate logic?

Now compare the semantics for O and $\mathsf{O}(\cdot/\cdot)$ from section 6.3.

$$\begin{array}{ll} M, w \models \mathsf{O}\, A & \text{iff } M, v \models A \text{ for all } v \in Min^{\prec_w}(\{u : wRv\}) \\ M, w \models \mathsf{O}(B/A) & \text{iff } M, v \models B \text{ for all } v \in Min^{\prec_w}(\{u : wRv \text{ and } M, v \models A\}) \end{array}$$

Informally, $\mathsf{O}\, A$ means that among the circumstantially accessible world, all the best worlds are A-worlds, while $\mathsf{O}(A/B)$ means that among the circumstantially accessible worlds *that are also* B*-worlds*, all the best worlds are A-worlds. The second argument place of $\mathsf{O}(\cdot/\cdot)$ restricts the domain of worlds over which O quantifies. As (2)–(4) illustrate, there is good reason to think that 'if'-clauses in English can play this role.

The upshot is that when a sentence appears to contain a conditional and a modal operator, like (1) or (3) or (4), the 'if'-clause sometimes serves to restrict the operator's domain.

Angelika Kratzer forcefully argued that this is what if-clauses *always* do:

> The history of the conditional is the story of a syntactic mistake. There is no two-place *if. . . then* connective in the logical forms of natural languages. *If*-clauses are devices for restricting the domains of various operators. Whenever there is no explicit operator, we have to posit one. [Kratzer 1991]

Recall the following example from section 8.2.

(5) If we leave after 5, we will miss the train.

Here there doesn't seem to be any modal operator. According to Kratzer, we therefore have to posit a hidden operator. Kratzer would suggest that the sentence contains a hidden epistemic necessity operator akin to 'must'. The if-clause restricts the domain of that operator.

So (5) is true iff *among the epistemically accessible worlds at which we leave after 5*, all worlds are worlds at which we miss the train. This is equivalent to saying that among the epistemically accessible worlds, all worlds are either worlds at which we don't leave after 5 or they are worlds at which we miss the train. So the strict epistemic translation $\Box(p \rightarrow q)$ gets the truth-conditions right, although Kratzer would insist that (5) is not composed of a one-place operator and a material conditional.

The case of subjunctive conditionals is similar. Here Kratzer might suggest that the 'if'-clause restricts a necessity operator with a non-epistemic, circumstantial flavour. Again, the resulting truth-conditions appear to be equivalent to those of the corresponding strict conditionals $\Box(A \rightarrow B)$.

However, this equivalence depends on the interpretation of the modal operator that is restricted by the 'if'-clause. Suppose we introduce an operator $\Box$ so that $\Box A$ is true iff A is true at the *closest* of the accessible worlds (relative to some suitable ordering). This is just how $\bigcirc$ works, except that the "closeness" ordering represents betterness. Restricting the domain of this $\Box$ operator effectively leads to the truth-conditions for variably strict conditionals.

The restrictor analysis of conditionals therefore does not determine a new logic of conditionals. Logically, 'if A then B' may well behave just like $A \strictif B$ or $A \boxright B$ (or even $A \rightarrow B$). Nonetheless, the analysis is worth knowing, because it sheds light on many otherwise puzzling fact about conditionals in natural language – for example,

- that 'if A then B' should sometimes be translated as '$A \wedge B$' (as in example (3)),
- that 'if A then it must be that B' should almost always be translated as $\Box(A \rightarrow B)$ – or $\Box(B/A)$ – even though 'must' appears to be in the consequent, and
- that sometimes, what appear to be combinations of modal operators and conditionals are best translated in terms of a primitive two-place operator (as in example (4)).

9 Modal Predicate Logic

9.1 Predicate logic recap

We are now going to add modal operators to the language of predicate logic. But first, let's briefly review the syntax and semantics of classical, non-modal predicate logic.

The language $\mathfrak{L}_P$ of predicate logic is made up of *predicates* $F, G, H, \ldots$, *(individual) constants* (or *names*) $a, b, c, \ldots$, *(individual) variables* $x, y, z, \ldots$, the logical symbols $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, $\forall$, $\exists$, and the parentheses (and). Individual variables and constants are also called *(singular) terms*.

Atomic sentences are formed by conjoining predicates with terms. Every predicate can only be applied to a fixed number of terms. Predicates that apply to one term are called *one-place* predicates; predicates that apply to two terms are called *two-place*, and so on. In English, for example, 'is human' is a one-place predicate because it combines with a single name to form a sentence ('Andy is human'), whereas 'loves' is a two-place predicate. In what follows, I will assume that F and G are one-place predicates while H is two-place. So Fa, Gy, and Hxa are sentences, but Fab and Hx are not.

It is often useful to add a special *identity predicate* '=' to the language of predicate logic. '=' is usually put between its two arguments; so we write '$a = b$' (sometimes in parentheses) instead of '$= ab$'. The identity predicate counts as a logical symbol because its meaning is held fixed across models: $a = b$ always means that a is the very same thing as b.

From atomic sentences, other sentences can be formed in the usual way by means of the truth-functional operators $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$. For example, since Fa and Gy are sentences, so are $(Fa \vee Gy)$ and $((Fa \wedge \neg Gy) \rightarrow Gy)$.

Another way to construct a new sentence from an old sentence is to add a quantifier in front. A *quantifier* is an expression of the form $\forall x$ or $\exists x$, where x is some variable. So if A is any sentence and x is a variable, then $\forall xA$ and $\exists xA$ are sentences. In $\forall xA$, the subsentence A and all its parts are said to be in the *scope* of the quantifier $\forall x$,

which is said to *bind* the variable x. If an occurrence of a variable lies in the scope of a quantifier which binds the variable, then the occurrence is called *bound*, otherwise it is *free*. For example, in the sentence $Fx \rightarrow \forall xFx$, the first occurrence of x is free and the last is bound.

So much for the syntax of predicate logic. Now for the semantics.

Like sentences of modal propositional logic, sentences of predicate logic are evaluated relative to a model. A model of predicate logic first of all specifies an *(individual) domain* D over which the quantifiers are said to range. $\exists xFx$ will be true if some individual in the domain has the property expressed by F. The second part of a model is an *interpretation function* V that assigns

(a) to each name a member of the domain D,

(b) to each one-place predicate a subset of D, and

(c) to each n-place predicate (for $n > 1$) a set of n-tuples from D.

An "n-tuple from D" is simply a list of length n, all elements of which are in D. Repetitions are allowed, so if d is a member of D, then $\langle d, d \rangle$ counts as a 2-tuple from D. 2-tuples are more commonly called *pairs*. We can subsume condition (b) under condition (c) by assuming that a 1-tuple from D is a member of D.

Informally, a model represents a conceivable scenario together with an interpretation of the non-logical vocabulary. In any scenario, there will be some things ("individuals") we want to talk about; these are represented by the domain. The non-logical vocabulary of predicate logic are the names and the predicates (except for the identity predicate). When we evaluate sentences at a scenario, the names $a, b, \ldots$ will pick out individuals in the scenario; so they are interpreted as denoting members of the domain. Predicates $F, G, \ldots$ express properties and relations among the relevant individuals. To determine the truth-value of, say Fa or $\exists xFx$ in a scenario, we only need to know which individuals in the scenario have the property expressed by F. Similarly, to determine the truth-value of Hab or $\exists x\exists yHxy$, we only need to know which pairs of individuals stand in the relation expressed by H. That's why an interpretation function in a model of predicate logic simply assigns sets of individuals or n-tuples of individuals to predicates. Fa is true in a given model iff the individual assigned to a in the model is a member of the set assigned to F; that is, iff $V(a) \in V(F)$.

Now we would like to specify how the truth-value of complex sentences in a model is determined by the meaning of their parts. For a quantified sentence like $\forall xFx$,

the immediate parts would be the quantifier $\forall x$ and the embedded sentence Fx. But what is the meaning of Fx? Interpretation functions assign individuals to names; they say nothing about variables. Even if we changed this and said that x should also be assigned a member of the domain, we would have to ignore the assignment when we interpret $\forall xFx$. We want $\forall xFx$ to be true iff Fx is true *no matter which individual is assigned to* x. We therefore define truth not just relative to a model, but relative to a model *and an assignment of individuals to variables*.

To illustrate, consider a model with just two individuals, denoted by a and b. Let $V(F)$ be the set that only contains the first individual, the one denoted by a. So Fa is true and Fb is false. The sentence Fx is neither true nor false, because the variable x does not pick out any particular individual. All we can say is that Fx is *true of* the first individual and not of the second, meaning that Fx comes out true if we assign the first individual to x and false if we assign the second individual to x. More generally, an $\mathfrak{L}_P$-sentence with free variables will typically be true relative to some assignment of values to the variables and false relative to others. $\forall xFx$ is true iff Fx is true relative to all assignments of values to the variable x.

In the formal semantics for $\mathfrak{L}_P$, truth is defined relative to a model $M = \langle D, V \rangle$ and a *variable assignment* g, understood as a function that maps variables to members of D. When we have nested quantifiers, as in $\forall x\exists yHxy$, we need to consider variable assignments that differ from other assignments with respect to a particular variable. $\forall x\exists yHxy$ is true iff, no matter what individual is assigned to x, there is some assignment of an individual to y that makes Hxy true. Equivalently: $\forall x\exists yHxy$ is true iff, for every variable assignment g, there is some variable assignment g' that differs from g at most in what it assigns to y, relative to which Hxy is true.

Let's say that (for any variable x) a variable assignment g' is an *x-variant* of a variable assignment g iff g' differs from g at most in the value it assigns to x.

Finally, let's introduce $[t]^{M,g}$ as shorthand for the individual picked out by the term t in model M relative to assignment t:

$$[t]^{M,g} =_{\text{def}} \begin{cases} \text{V(t)} & \text{if } t \text{ is a name (and } V \text{ is the interpretation function of } M) \\ \text{g(t)} & \text{if } t \text{ is a variable.} \end{cases}$$

We then have all the ingredients to state the standard semantics of classical predicate logic.

Definition 9.1: Semantics of predicate logic

If $M = \langle D, V \rangle$ is an $\mathfrak{L}_P$-model, ϕ is an n-place predicate (for $n \geq 1$), $s, t, t_1, \ldots, t_n$ are terms, and g is a variable assignment, then

(a)	$M, g \models \phi t_1 \ldots t_n$	iff $\langle [t_1]^{M,g}, \ldots, [t_n]^{M,g} \rangle \in V(\phi)$.
(b)	$M, g \models s = t$	iff $[s]^{M,g} = [t]^{M,g}$.
(c)	$M, g \models \neg A$	iff $M, g \not\models A$.
(d)	$M, g \models A \wedge B$	iff $M, g \models A$ and $M, g \models B$.
(e)	$M, g \models A \vee B$	iff $M, g \models A$ or $M, g \models B$.
(f)	$M, g \models A \rightarrow B$	iff $M, g \models B$ or $M, g \not\models A$.
(g)	$M, g \models A \leftrightarrow B$	iff $M, g \models (A \rightarrow B)$ and $M, g \models (B \rightarrow A)$.
(h)	$M, g \models \forall x A$	iff $M, g' \models A$ for all x-variants g' of g.
(i)	$M, g \models \exists x A$	iff $M, g' \models A$ for some x-variant g' of g.

Definition 9.1 defines truth relative to a model and an assignment function. We can define *truth in a model* as truth relative to all assignment functions for the model. That is, A is true in M iff $M, g, \models A$ for every variable assignment g. A is *valid* iff A is true in all models.

With the help of definition 9.1, we could in principle work out whether various $\mathfrak{L}_P$-sentences are valid or invalid. But the process is tedious. As an easier method, we can use trees.

The tree rules for classical predicate logic look a lot like those for modal propositional logic, except that sentences are no longer relativised to worlds. We also need new rules for the quantifiers and the identity symbol. To motivate the quantifier rules, let's do an example. We'll try to prove $\forall x(Fx \wedge Gx) \rightarrow \forall x Fx$. We begin with the standard rule for negated conditionals:

1.	$\neg(\exists x(Fx \wedge Gx) \rightarrow \exists x Fx)$	(Ass.)
2.	$\exists x(Fx \wedge Gx)$	(1)
3.	$\neg \exists x Fx$	(1)

Node 2 says that $Fx \wedge Gx$ is true of some individual. To expand this node, we introduce a new name a for that individual, and infer $Fa \wedge Ga$.

4.	$Fa \wedge Ga$	(2)
5.	Fa	(4)
6.	Ga	(4)

Nodes 5 and 6 expand the conjunction on node 4. Next, we expand node 3, which says that Fx is true of nothing. It follows that Fa must be false:

7.	$\neg Fa$	(3)
	x	

The tree is closed because the sentence on node 7 is the negation of the sentence on node 5.

To state the general rules, we need some more notation. If A is a sentence, x is a variable, and c is a name, let $A[c/x]$ be the sentence obtained from A by replacing all free occurrences of x with c. So $Fx[a/x]$ is Fa, but $\forall xFx[a/x]$ is $\forall xFx$ because this sentence contains no free occurrences of x.

The general rule for expanding nodes of type $\exists xA$ states that $A[c/x]$ should be added to all open branches below the node, where c is a "new" name – one that does not already occur on the relevant branch. The original $\exists xA$ node can then be ticked off. $\forall xA$ nodes can be expanded several times, once for each "old" name. So if $\forall xA$ occurs on a branch, and the branch contains the names a and b, we can add both $A[a/x]$ and $A[b/x]$. We should not expand $\forall xA$ nodes with a new name except if no name occurs on a branch. $\forall xA$ nodes are never ticked off.

Here is a summary of the quantifier rules; 'old or first' means that the relevant name either already occurs on the branch or it is introduced as the first name on the branch.

$\forall xA$	$\exists xA$	$\neg\forall xA$	$\neg\exists xA$
⋮	⋮	⋮	⋮
$A[c/x]$	$A[c/x]$	$\neg A[c/x]$	$\neg A[c/x]$
↑	↑	↑	↑
old or first	new	new	old or first

If we have the identity predicate, we need two more rules. First, for any name c that occurs on a branch, we may at any point add the node $c = c$. The second rule is

known as **Leibniz' Law**. Informally, Leibniz' Law states that if something is true of a, and a is (numerically) identical to b, then the same thing is true of b. In the tree method, Leibniz' Law takes the following form: if an identity statement $b = c$ occurs on a branch, as well as some sentence A which contains b, then we may add a new node with the same sentence A except that one or more occurrences of b have been replaced by c, or one or more occurrences of c by b. Let $A[c//b]$ stand for any sentence that results from A by replacing one or more occurrences of b with c. The identity rules can then be summarized as follows.

Identity	Leibniz' Law	Leibniz' Law
⋮	$b = c$	$b = c$
$c = c$	A	A
↑	⋮	⋮
old	$A[c//b]$	$A[b//c]$

The following tree uses Leibniz' Law to prove $(Haa \wedge a=b) \rightarrow Hab$.

1.	$\neg((Haa \wedge a=b) \rightarrow Hab)$	(Ass.)
2.	$Haa \wedge a=b$	(1)
3.	$\neg Hab$	(1)
4.	Haa	(2)
5.	$a=b$	(2)
6.	Hab	(4, 5, LL)
	x	

Exercise 9.1

Give tree proofs for the following sentences.

(a) $\forall x Fx \rightarrow Fa$

(b) $\exists x \forall y Hxy \rightarrow \forall y \exists x Hxy$

(c) $\exists y \forall x (Fy \to Fx)$
(d) $\forall x (x = x)$
(e) $\forall x \forall y (x = y \to y = x)$

Exercise 9.2

Show that the second version of the Leibniz' Law rule can be derived from the first.

There is also an axiomatic calculus for predicate logic. In fact, there are many sound and complete calculi. A simple one extends the axiomatic calculus of propositional logic by the following axiom schemas, where A is any sentence, x any variable, and b, c are any names:

$$\forall x A \to A[c/x] \qquad \textbf{(UI)}$$
$$\forall x (x = x) \qquad \textbf{(ID)}$$
$$b = c \to (A \to A[c//b]) \qquad \textbf{(LL)}$$

'**UI**' stands for *Universal Instantiation*; **ID** and **LL** are axiomatic versions of the basic identity rule and Leibniz' Law. We also need one new rule, in addition to *Modus Ponens*, a form of *Universal Generalisation*:

If $A \to B[c/x]$ occurs on a proof, and x is not free in A,
then $A \to \forall x B$ may be added. **(UG)**

9.2 De dicto and de re

Now let's add the (monadic) sentence operators $\Box$ and $\Diamond$ to $\mathfrak{L}_P$. The result is the basic language of modal predicate logic, $\mathfrak{L}_{MP}$. For some purposes, several boxes and diamonds will be needed, but let's focus on mono-modal predicate logics. As in earlier chapters, I won't fix the interpretation of the modal operators. The box might formalize physical necessity, metaphysical necessity, provability, knowledge, obligation, "it is always going to be the case", or various other concepts.

In the language of modal predicate logic, we can say things like $\Box \exists x Fx$ and

$\exists x \Box F x$. Imagine a lottery. If we read the box as 'it is certain that', and F as 'winning', then $\Box \exists x F x$ translates 'it is certain that someone is winning'. By contrast, $\exists x \Box F x$ translates 'there is someone of whom it is certain that they are winning'. We say that the second sentence is **de re**, because it assert *of* a particular *thing* (Latin, *de re*) that it has a modal property – the property of being the certain winner. The first sentence, $\Box \exists x F x$, merely states that the proposition (Latin, *dictum*) $\exists x F x$ is certain. Sentences like this are called **de dicto**.

In general, a sentence is *de re* if it contains a modal operator in whose scope lies a name or a variable that is not bound within the scope of the operator. For example, $\Diamond F a$, $\forall x(F x \to \Box G x)$ and $\Box(\forall x F x \to F a)$ are *de re*, but $\Diamond \forall x F x \to F a$ is not. Any sentence that contains a modal operator and is not *de re* is *de dicto*.

In English, many sentences are ambiguous between a *de re* reading and a *de dicto* reading. For example, 'something necessarily exists' can mean either that there is an object which could not have failed to exist, or that it could not have been the case that nothing exists. The first reading is *de re*, the second *de dicto*.

Exercise 9.3

Translate the following sentences into modal predicate logic. (Some of them are ambiguous.)

(a) John must be hungry.
(b) Every cyclist must have legs.
(c) Every day might be our last.
(d) If water is H_2O then cucumbers might contain H_2O.
(e) If anyone has to leave early, they should do so quietly.
(f) Everyone who bought a ticket is allowed to enter.
(g) Alice knows that there is someone who knows that she has a lover.
(h) One day, all those who are rich will be poor.

Exercise 9.4

Which of your translations from the previous exercise are *de re* and which are *de dicto*?

The complexity of modal predicate logic largely arises from questions about the

treatment of *de re* sentences. Indeed, on some interpretations of the modal operators, it is questionable whether one can even make sense of *de re* modality. For example, suppose we interpret the box as 'it is analytic that' or 'it is provable that'. The things that are analytic or provable are, in the first place, sentences. The sentence '$\forall x Fx \to Fa$', for example, is provable in classical predicate logic, and 'all vixens are female foxes' is analytic in English. It is not clear what it could mean to say that something is provable or analytic *of* a particular thing.

To illustrate, let's introduce the name 'Julius' for whoever invented the zip. The sentence 'Julius invented the zip (provided that someone invented the zip)' is then analytic. Can we infer that it is analytic *of* the person who invented the zip that they invented the zip? The problem is that this person has multiple names, and depending on which name we plug into the schema '— invented the zip', we sometimes get an analytic truth and sometimes not. For 'Julius', the sentence is analytic; for whatever name the inventor of the zip was given by their parents, the sentence is not analytic.

These worries were prominently raised by W.V.O. Quine in the 1940s. They have since faded, in part because philosophers have turned their attention to other interpretations of the box and the diamond for which the problem is less acute. For example, many philosophers believe that individuals have essential properties which they could not possibly fail to have, and which are independent of how the individuals are described or picked out. Saul Kripke argued that for humans, ancestry is essential: you could not have had other parents; anyone born from other parents would not have been you. If $\Box A$ means that A could have been the case, then on this view sentences like $\Box Fa$ or $\exists x \Box Fx$ are clearly intelligible.

In any case, *de re* sentences frequently come up in ordinary and philosophical discourse. Let's see what problems they create, apart from Quine's worries about intelligibility.

9.3 Existence

We can easily combine proof methods for classical predicate logic with methods for modal propositional logic. For example, we could get a proof method for modal predicate logic by adding schema **K** and the rule of necessitation to the axiomatic calculus for predicate logic. Equivalently, we could add the K-rules for the box and the diamond to the tree rules for predicate logic (and add world labels to the

latter rules). We then get some interesting results. For example, we can prove the following.

$$\forall x \Box \exists y\, (y = x) \qquad \textbf{(NE)}$$

Here is a tree proof.

1.	$\neg \forall x \Box \exists y\, (y = x)$	(w)	(Ass.)
2.	$\neg \Box \exists y\, (y = a)$	(w)	(1)
3.	wRv		(2)
4.	$\neg \exists y\, (y = a)$	(v)	(2)
5.	$\neg (a = a)$	(v)	(4)
6.	$a = a$	(v)	(Id.)
	x		

If we read the box as 'necessarily', then **NE** says that for every thing x there necessarily exists some thing y which is identical to x. But if there exists some thing y which is identical to x, then x itself exists. Conversely, if x exists, then there exists some thing y (namely, x) which is identical to x. So we can understand $\exists y\, (y = x)$ as saying that x exists. And so **NE** states that everything exists necessarily. (Hence 'NE', for 'necessary existence'.)

We can reach essentially the same conclusion even without the identity rules. The necessity of existence then can't be expressed by a sentence, but it can be expressed by the following schema, known as the **Converse Barcan Formula**:

$$\Box \forall x A \rightarrow \forall x \Box A \qquad \textbf{(CBF)}$$

We will see precisely how **CBF** corresponds to the necessity of existence later, when we have introduced suitable models, but the point is not hard to see. Let F stand for a property which everything has at all and only the worlds at which it exists (such as the property of existing). Then $\Box \forall x F x$ is true at all worlds. If **CBF** is valid, it follows that $\forall x \Box F x$ is true as well. But $\forall x \Box F x$ is false at any world at which something exists that fails to exist at some accessible world. So if **CBF** is valid, there can be no such worlds.

Below is a (schematic) proof of **CBF** in the combined axiomatic calculus for predicate logic and the modal logic K.

1.	$\forall x A \rightarrow A[c/x]$	(UI)
2.	$\Box(\forall x A \rightarrow A[c/x])$	(from 1 by Nec)
3.	$\Box(\forall x A \rightarrow A[c/x]) \rightarrow (\Box\forall x A \rightarrow \Box A[c/x])$	(K)
4.	$\Box\forall x A \rightarrow \Box A[c/x]$	(from 2 and 3 by MP)
5.	$\Box\forall x A \rightarrow \forall x \Box A$	(from 4 by UG).

The Converse of the Converse Barcan Formula is the original **Barcan Formula**, first discussed (and defended) by Ruth Barcan Marcus in 1946:

$$\forall x \Box A \rightarrow \Box \forall x A \qquad \textbf{(BF)}$$

BF cannot be derived in the combined proof systems for classical predicate logic and K, but it can be derived if the modal basis is strengthened to B or S5.

Exercise 9.5

Give a tree proof of **BF** using the S5 rules and the rules of predicate logic.

While **CBF** effectively rules out the possibility of actual things failing to exist at other accessible worlds, **BF** rules out the possibility of things at other worlds failing to exist at the actual world. Here's why. Let F be a property which all things that exist at the actual world have at all accessible worlds. If **BF** is valid, it follows that all things at those worlds also have F. But if the worlds had extra individuals, these individuals could fail to have F.

What shall we make of these results? Three main answers have been pursued.

First, we might simply accept the results. Some philosophers have argued that indeed, nothing that exists could have failed to exist, and nothing could have existed that doesn't actually exist. On this metaphysical picture, **NE** is true, and so are all instances of **BF** and **CBF**, if the box is read as a suitable kind of metaphysical necessity. It is still a little odd to think that such a substantive metaphysical doctrine could be established by pure logic. Moreover, the problem raised by **NE**, **BF**, and **CBF** doesn't just affect metaphysical modality. In epistemic logic, we may not want

to say that everything that exists is known to exist, or that whenever someone can't rule out that something exists, then that thing actually exists. In temporal logic, we may not want to say that everything that currently exists is always going to exist: there is no logical guarantee that the Arctic ice shield will still exist in 2050.

Another response is to revise the interpretation of the quantifiers. So far, I have assumed that the quantifiers range only over things that exist at the actual world (or at the present time, in temporal logic). This is a common assumption, but $\mathfrak{L}_{MP}$ is a made-up language, so we can make its expressions mean whatever we want. In particular, we could take the quantifiers to range over all things that exist at any possible world. We would thereby re-interpret $\exists$ to mean 'there could have been' rather than 'there is'. On this reading, **NE**, **BF**, and **CBF** are plausibly harmless. For example, **NE** no longer asserts the necessity of existence. Rather, it states that for all possible things y it is necessary that there is some possible thing x which is identical to y.

Conceptually, this second strategy has the downside of abandoning the core idea of modal logic that sentences are to be evaluated *locally*, relative to a particular world, with only the modal operators providing access to what happens at other worlds.

A third response is to find a fault in the proofs of **NE**, **BF**, and **CBF**. To be sure, the proofs only use familiar rules from predicate logic and modal propositional logic, but perhaps these rules cannot simply be combined.

To figure out which rules of proof are acceptable and which aren't, we need a better grip on validity and logical consequence in modal predicate logic. So let's switch from the proof-theoretic perspective to the model-theoretic perspective.

Exercise 9.6

Consider the following four schemas.

(1) $\Diamond\exists x A \rightarrow \exists x \Diamond A$
(2) $\Box\exists x A \rightarrow \exists x \Box A$
(3) $\exists x \Box A \rightarrow \Box\exists x A$
(4) $\exists x \Diamond A \rightarrow \Diamond\exists x A$

(a) Which of (1)–(4) are equivalent to **BF** or **CBF**, given the duality of $\Box$ and $\Diamond$ and of $\forall x$ and $\exists x$?
(b) Are the other schemas provable as well? Are they plausible?

Exercise 9.7

The **K**-like principle $\forall x(A \to B) \to (\forall xA \to \forall xB)$ is easily provable in classical predicate logic. Can you see how the strict version $\forall x(A \mathrel{-\!\!\!\exists} B) \to (\forall xA \mathrel{-\!\!\!\exists} \forall xB)$ is related to the Barcan Formula?

9.4 Constant domain semantics

As always, a model of modal predicate logic is meant to represent a conceivable scenario together with an interpretation of the non-logical vocabulary. For $\mathfrak{L}_{MP}$, the non-logical vocabulary consists of the names and predicates from $\mathfrak{L}_P$. Like in modal propositional logic, we will assume that a relevant scenario involves a plurality of "worlds" relative to which sentences are interpreted as true or false; the modal operators function as quantifiers over accessible worlds.

In chapter 2 I explained that a world is commonly understood as a maximally specific way things might have been. That's why every sentence letter of modal propositional logic has a determinate truth-value at every world, no matter how it is interpreted (provided the interpretation is unambiguous and precise). For the purposes of computing the truth-values of complex $\mathfrak{L}_M$-sentences, these world-relative truth-values were all we needed to know about the meaning of the sentence letters. In $\mathfrak{L}_{MP}$, we instead need to know something about the meaning of the names and predicates, and about the domain over which individual quantifiers are meant to range.

Let's begin be assuming that the quantifiers range over the same domain of individuals relative to every world. As I mentioned in the last section, this could be motivated either by the assumption that the very same things exist at all worlds, or by re-interpreting the quantifiers to range over all possible individuals. Every name, we will assume, picks out some individual in the domain. Predicates still express properties and relations, but these can no longer be represented as sets or tuples from the domain. That's because, whether a given individual has a given property often varies from world to world. In this world I'm having tea, in others I'm having coffee. So if F expresses the property of having tea, then the set of individuals to whom F applies varies from world to world. To determine the truth-value of arbitrary $\mathfrak{L}_{MP}$-sentences at arbitrary worlds, we need to know to which individuals, or tuples of individuals, a predicate applies at any world.

These considerations lead to the following definition of a model.

Definition 9.2

A **constant domain model** for modal predicate logic is a structure M consisting of

1. a non-empty set of "worlds" W,
2. an accessibility relation R on W,
3. a non-empty set D (the individual domain), and
4. an interpretation function V that assigns
   - to each name a member of D, and
   - to each n-place predicate and world w in W a set of n-tuples from D.

Truth of $\mathfrak{L}_{MP}$-sentence at worlds in constant domain models is defined as follows.

Definition 9.3: Constant domain semantics

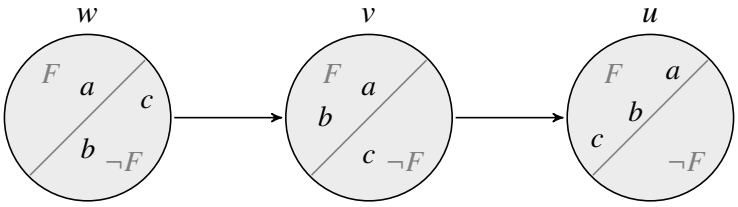
If $M = \langle W, R, D, V \rangle$ is a constant domain model, w is a member of W, ϕ is an n-place predicate (for $n \geq 1$), $s, t, t_1, \ldots, t_n$ are terms, and g is a variable assignment, then

(a)	$M, w, g \models \phi t_1 \ldots t_n$	iff $\langle [t_1]^{M,g}, \ldots, [t_n]^{M,g} \rangle \in V(\phi, w)$.
(b)	$M, w, g \models s = t$	iff $[s]^{M,g} = [t]^{M,g}$.
(c)	$M, w, g \models \neg A$	iff $M, w, g \not\models A$.
(d)	$M, w, g \models A \wedge B$	iff $M, w, g \models A$ and $M, w, g \models B$.
(e)	$M, w, g \models A \vee B$	iff $M, w, g \models A$ or $M, w, g \models B$.
(f)	$M, w, g \models A \rightarrow B$	iff $M, w, g \models B$ or $M, w, g \not\models A$.
(g)	$M, w, g \models A \leftrightarrow B$	iff $M, w, g \models (A \rightarrow B)$ and $M, w, g \models (B \rightarrow A)$.
(h)	$M, w, g \models \forall x A$	iff $M, w, g' \models A$ for all x-variants g' of g.
(i)	$M, w, g \models \exists x A$	iff $M, w, g' \models A$ for some x-variant g' of g.
(j)	$M, w, g \models \Box A$	iff $M, v, g \models A$ for all $v \in W$ such that wRv.
(k)	$M, w, g \models \Diamond A$	iff $M, v, g \models A$ for some $v \in W$ such that wRv.

Moreover, for any sentence A,

(l) $M, w \models A$ iff $M, w, g \models A$ for all variable assignments g.

As an example, here is a (partial) picture of a constant domain model, with three worlds and three individuals.



At world w, only the individual denoted by a has the property expressed by F; at world v, both a and b have F, and at u, everything is F. So Fa is true at w. Fx is not true at w, because it is false relative to an assignment g that maps x to the individual b. But Fx is true at w relative to a different assignment g that maps x to a. By definition 9.3, this means that $\exists xFx$ is true at w. Along the same lines, we can figure out that $\exists xFx$ is true at v. Since w can see v, it follows that $\Diamond\exists xFx$ is true at w. The *de re* sentence $\exists x\Diamond Fx$ is also true at w, because $\Diamond Fx$ is true at w relative to an assignment g that maps g to (say) b.

Exercise 9.8

Which of the following statements are true for the above model M?

(a) $M, w \models \Diamond Fc$
(b) $M, w \models \Box\exists xFx$
(c) $M, w \models \forall x\Box Fx \rightarrow \Box\forall xFx$
(d) $M, v \models \neg\forall x\Box\neg Fx$
(e) $M, u \models \Box\forall xFx$
(f) $M, w \models \forall x(Fx \leftrightarrow \Box Fx)$
(g) $M, w \models \Box\forall x(\Diamond Fx \rightarrow Fx)$

Let's say that a sentence is **CK-valid** if it is true at all worlds in all constant domain models. A schema is CK-valid if all its instances are CK-valid.

It is not hard to show that **NE**, **BF**, and **CBF** are all CK-valid. I'll go through the argument for the Barcan Formula.

Observation 9.1: **BF** is CK-valid.

Proof. Suppose (some instance of) $\forall x \Box A$ is true at some world w in some constant domain model relative to some assignment g. By clause (h) of definition 9.3, it follows that $\Box A$ is true at w relative to every x-alternative g' of g. By clause (j) of definition 9.3, it follows that A is true at every world v accessibility from w relative to every x-alternative g' of g. By clause (h), this means that $\forall x A$ is true relative to g at every world v accessible from w. So by clause (j), $\Box \forall x A$ is true at w relative to g. This shows that whenever $\forall x \Box A$ is true at some world w relative some assignment g, then $\Box A \forall x A$ is also true at w relative to g. By clause (f) of definition 9.3, it follows that $\forall x \Box A \rightarrow \Box A \forall x A$ is also true at every world relative to every assignment. □

Earlier I mentioned that the Barcan Formula is not provable by combining the (axiomatic or tree) methods for classical predicate logic with the methods for the modal logic K. So while the combined methods are *sound* with respect to CK-validity – everything that is provable is CK-valid – they are not *complete*, because not every CK-valid sentence is provable. Instances of the Barcan Formula are one exception. Another exception is the "necessity of distinctness" formula **ND**.

$$\forall x \forall y (\neg (x = y) \rightarrow \Box \neg (x = y)) \qquad \textbf{(ND)}$$

Exercise 9.9

Show that **ND** is CK-valid.

If we merge the axiomatic calculus for K and the axiomatic calculus for predicate logic, and add **BF** and **ND** as further axioms, we get a calculus that is sound and complete with respect to constant domain semantics: all and only the CK-valid sentences are provable. Completeness can be established by adapting the canonical model method; we won't go through the details.

Further axioms are required if we add constraints on the accessibility relation. For example, to get a sound and complete calculus for the sentences that are true at all worlds in all *reflexive* constant domain models, the **T**-schema must be added. Similarly, the **4**-schema must be added if we require the accessibility relation to be transitive, and the **5**-schema if we require it to be euclidean.

In general, the correspondence facts from chapter 3 carry over to modal predicate logic with a constant domain semantics: **T** corresponds to reflexivity, **4** to transitivity, and so on. But completeness facts do not generally carry over. For example, you might expect that adding the schemas **T**, **4**, and **G** (giving us the quantified analogue of S4.2) would result in a complete calculus for reflexive, transitive, and convergent models. But it does not. The resulting calculus is not complete with respect to any class of constant domain models.

9.5 Variable domain semantics

In the previous section, we assumed that at each world, the quantifiers range over the very same individuals. Now let's relax this assumption and allow for variable (instead of constant) domains.

We then have to confront the following issue. Consider the sentence $\Box\exists x(x=a)$, which might be read as "a necessarily exists". The sentence is valid in constant domain semantics, but now we want it to be invalid, as we want to allow for models in which the individual denoted by a does not exist at every world. At worlds where the individual does not exist, $\exists x(x=a)$ will be false, and so $\Box\exists x(x=a)$ will be false at any world that has access to such a world.

The issue we have to confront is how, in general, we should evaluate a sentence at a world if the sentence contains a name that doesn't refer to any individual at that world. If a does not exist, is Fa always false? How about $\neg Fa$, or $Fa \vee \neg Fa$?

The issue arises not just in modal logic. In classical predicate logic, we assume that every name picks out some individual. In ordinary language, however, there are names that don't pick out anything. In the mid-19th century it was observed that the movement of the planet Mercury conforms to Newton's Laws only if there is another planet on an orbit in between Mercury and the Sun. With great confidence in Newton's Laws, Astronomers baptised this as yet undiscovered planet 'Vulcan'. But Vulcan has never been found. The relevant movement of Mercury was eventually explained by Einstein's theory of general relativity. The name 'Vulcan' has turned out to be empty; it does not refer to anything. So we can't simply assume that all names refer. Even if we have good reason to think that a name refers, we might turn out to be wrong.

So what should we say about the truth-value of sentences containing an empty name? One possibility is to say that all such sentences are false, or neither true

nor false. But this would mean that even 'Vulcan does not exist' (or the predicate logic translation $\neg\exists x(x = a)$) is not true. Another possibility is to say that *atomic* sentences with empty names are always false. The idea is that atomic sentences attribute properties to individuals, and if something doesn't exist then it doesn't have any properties. A third (metaphysically more extravagant) option is to hold that even non-existent objects can have properties, so that empty names behave much like ordinary, non-empty names.

These options are formally investigated in a branch of logic called **free logic**. Instead of surveying all varieties of free logic, I will choose the second of the above options, which assumes that atomic sentences with empty names are always false. This gives rise to what is known as a **negative free logic**.

Negative free logic is weaker than classical predicate logic, because classically valid sentences like $\exists x(x = a)$ are no longer valid. Indeed, even $a = a$ is invalid in negative free logic, because it is an atomic sentence and hence taken to be false if a is an empty name. The classical axiom $\forall x(x = x)$, however, is valid: any thing x is identical to itself. It follows that the **UI** principle $\forall xA \rightarrow A[c/x]$ must be weakened. From $\forall xA$ we can only infer $A[c/x]$ if we also know that c exists (i.e., that the name c is not empty). **UI** is therefore replaced by the principle of "free universal instantiation", **FUI**:

$$\forall xA \rightarrow (\exists x(x = c) \rightarrow A[c/x]) \qquad \textbf{(FUI)}$$

In the tree method, the rule for $\forall xA$ is similarly replaced by a branching rule on which we put $\neg\exists(x = c)$ on one branch and $A[c/x]$ on the other.

If you look back at the proofs of **NE**, **BF**, and **CBF** in section 9.3, you can see that they all involved the original **UI** principle (or the corresponding tree rule). But that principle is invalid if names can be empty. And if we don't assume that whatever exists at one world also exists at every accessible world, then names *can* be empty, in the sense that we need to evaluate sentences at worlds at which the individual denoted by a name in the sentence does not exist.

Definition 9.4

A **variable domain model** for modal predicate logic is a structure M consisting of

1. a non-empty set of "worlds" W,

2. an accessibility relation R on W,
3. for each world w in W, a non-empty individual domain D_w, and
4. an interpretation function V that assigns
   - to each name a member of some domain D_w, and
   - to each n-place predicate and world w in W a set of n-tuples from D_w.

Truth of $\mathfrak{L}_{MP}$-sentence at worlds in variable domain models is defined as follows.

Definition 9.5: Variable domain semantics

If $M = \langle W, R, D, V \rangle$ is a variable domain model, w is a member of W, ϕ is an n-place predicate (for $n \geq 1$), $s, t, t_1, \ldots, t_n$ are terms, and g is a variable assignment, then

(a)	$M, w, g \models \phi t_1 \ldots t_n$	iff $\langle [t_1]^{M,g}, \ldots, [t_n]^{M,g} \rangle \in V(\phi, w)$.
(b)	$M, w, g \models s = t$	iff $[s]^{M,g} = [t]^{M,g}$ and $[s]^{M,g}$ and $[t]^{M,g}$ are in D_w.
(c)	$M, w, g \models \neg A$	iff $M, w, g \not\models A$.
(d)	$M, w, g \models A \wedge B$	iff $M, w, g \models A$ and $M, w, g \models B$.
(e)	$M, w, g \models A \vee B$	iff $M, w, g \models A$ or $M, w, g \models B$.
(f)	$M, w, g \models A \rightarrow B$	iff $M, w, g \models B$ or $M, w, g \not\models A$.
(g)	$M, w, g \models A \leftrightarrow B$	iff $M, w, g \models (A \rightarrow B)$ and $M, w, g \models (B \rightarrow A)$.
(h)	$M, w, g \models \forall x A$	iff $M, w, g' \models A$ for all x-variants g' of g for which $g'(x) \in D_w$.
(i)	$M, w, g \models \exists x A$	iff $M, w, g' \models A$ for some x-variant g' of g for which $g'(x) \in D_w$.
(j)	$M, w, g \models \Box A$	iff $M, v, g \models A$ for all $v \in W$ such that wRv.
(k)	$M, w, g \models \Diamond A$	iff $M, v, g \models A$ for some $v \in W$ such that wRv.

Moreover, for any sentence A,

(l) $M, w \models A$ iff $M, w, g \models A$ for all variable assignments g.

An adequate proof system for variable domain semantics requires some more adjustments than replacing **UI** by **FUI**. For example, the Identity rule for trees must be changed so that $c = c$ can only be added to a branch if the branch contains

$\exists x(x = c)$. Completeness can still be proved with canonical models, but the proofs are considerably more complicated than in constant domain semantics.

It is easy to check that **NE**, **CBF**, and **CBF** are all invalid in variable domain semantics. To render **NE** and **CBF** valid, we would have to assume that whenever a world w has access to a world v, then all individuals at w also exist at v. For **BF**, we would conversely have to assume that all individuals at v also exist at w.

More precisely, let a **variable domain frame** be a variable domain model without the interpretation function V, and let's say that a sentence or schema is **valid on a variable domain frame** iff it is true at all worlds in all models based on that frame. (Compare section 3.4.) A frame has **decreasing domains** if, whenever wRv then $D_v \subseteq D_w$; a frame has **increasing domains** if, whenever wRv then $D_w \subseteq D_v$. Now, **CBF** and **NE** both correspond to increasing domains, while **BF** corresponds to decreasing domains:

Observation 9.2:

(i) **CBF** is valid on a variable domain frame iff the frame has increasing domains.

(ii) **NE** is valid on a variable domain frame iff the frame has increasing domains.

(iii) **BF** is valid on a variable domain frame iff the frame has decreasing domains.

Proof of (i). Suppose a frame F does not have increasing domains. Then F contains a world w in whose domain D_w lies an individual d which does not exist at some accessible world v. Let V be an interpretation function so that $V(F, w) = D_w$ and $V(F, v) = D_v$. In the resulting model, $\Box\forall xFx$ is true at w, but $\forall x\Box Fx$ is false, since d is not in $V(F, v)$. So **CBF** is not true in all models based on frame F.

In the other direction, suppose **CBF** is not valid in a frame F. Then there is a world w in some model M based on F at which some instance of $\Box\forall xA$ is true while $\forall x\Box A$ is false. If $\forall x\Box A$ is false at w, then there is some w-accessible world v at which A is false for some individual d in D_w. But since $\Box\forall xA$ is true at w, A is true of all members of D_v. So d is not in D_v. And so F does not have increasing domains.

The proofs of (ii) and (iii) are analogous. □

Exercise 9.10

Definition 9.4 requires every name in every model to pick out a possible individual. In that sense, it does not allow for genuinely empty names. How could we change definitions 9.4 and 9.5 if we wanted to allow for names that don't pick out possible individuals?

9.6 Contingent identity

In exercise 9.4 (page 182) you showed that the "necessity of distinctness" is valid in constant-domain semantics.

$$\forall x \forall y(\neg(x = y) \rightarrow \Box\neg(x = y)) \qquad \textbf{(ND)}$$

The same is true for the "necessity of identity", **NI**:

$$\forall x \forall y(x = y \rightarrow \Box(x = y)) \qquad \textbf{(NI)}$$

I didn't mention this at the time, because unlike **ND**, **NI** is provable with the combined resources of predicate logic and the modal logic K. I give a tree proof.

1.	$\neg\forall x \forall y(x = y \rightarrow \Box(x = y))$	(w)	(Ass.)
2.	$\neg\forall y(a = y \rightarrow \Box(a = y))$	(w)	(1)
3.	$\neg(a = b \rightarrow \Box(a = b))$	(w)	(2)
4.	$a = b$	(w)	(3)
5.	$\neg\Box(a = b)$	(w)	(3)
6.	$\neg\Box(b = b)$	(w)	(5, LL)
7.	wRv		(6)
8.	$\neg(b = b)$	(v)	(8)
9.	$b = b$	(v)	(Id.)
	x		

Exercise 9.11

Give a tree proof of **ND** using the combined rules of classical predicate logic and the *S5* rules.

In the variable-domain semantics from the previous section, **ND** is still valid. **NI** becomes invalid, because identity statements involving non-existent individuals are false. However, the following weakened principle remains valid:

$$\forall x \forall y (x = y \rightarrow \Box(\exists z(z = x) \rightarrow x = y)) \qquad \textbf{(WNI)}$$

This says that whenever x is identical to y then x is identical to y at all accessible worlds *at which x exists*.

Like the necessity of existence, the necessity of identity and distinctness has given rise to some controversy. After all, it may seem obvious that identity facts are often contingent. For example, Donald Trump is identical to the 45th President of the United States. But it could easily have been otherwise. If Hillary Clinton had won the electoral college, then she would have been the 45th President. So even though Clinton is not identical to the 45th President, she could have been. And even though Trump is identical to the 45th President, he could have failed to be (and without failing to exist). So we seem to have counterexamples to the following principles, which are entailed by **ND** and **WNI**:

$$\neg(a = b) \rightarrow \Box\neg(a = b) \qquad \textbf{(ND′)}$$

$$a = b \rightarrow (\exists z(z = a) \rightarrow \Box(a = b)) \qquad \textbf{(WNI′)}$$

Again, there are several possible lines of response. One is to accept the necessity of all identity statements. Perhaps we are mistaken when we think that Clinton could have won, or that anyone could fail to know that Trump is the 45th President?

Another response is to diagnose a fault in our semantics, whose fix would explain where the proof of **NI** (and **ND**, in exercise 9.6) went awry. An obvious culprit is our account of names. In both constant-domain and variable-domain models, we have assumed that the interpretation function simply assigns an individual to every name. A modal sentence like $\Box Fa$ is then evaluated as true (at a world) iff the individual assigned to the name a has the property expressed by F at every accessible world. But if we translate ‘the 45th President’ as a name, then this name will pick out

different people at different worlds. In the actual world, it picks out Donald Trump, in other worlds Hillary Clinton.

Terms that pick out the same individual at every possible world are called **rigid**. The semantics we have given for $\mathfrak{L}_{MP}$ treats names as rigid. But descriptions like 'the 45th President' are non-rigid.

So perhaps we should change our semantics to allow for non-rigid names. I will explain how this could be done later. First I want to mention a third line of response.

According to the third response, it is wrong to translate descriptions like 'the 45th President of the US' as names. The obvious counterexamples to **ND′** and **WNI′** would therefore rest on a faulty translation.

Bertrand Russell already pointed out in 1905 that treating descriptions as names leads to a host of problems even in non-modal predicate logic. For example, since there is no king of France, 'the king of France is bald' and 'the king of France is not bald' both appear to be false. If we translate 'the king of France' as a name, we would have to conclude that Fa and $\neg Fa$ are both false.

Russell argued that 'the king of France is bald' is true iff (i) some king of France is bald, and (ii) there is no more than one king of France. Instead of translating the sentence as Fa, we should therefore translate it as follows:

$$\exists x(Fx \wedge \forall y(Fy \rightarrow x = y) \wedge Gx)$$

Russell's analysis predicts – correctly – that descriptions have scope. For example, 'the king of France is not bald' can mean either that there is a king of France who is not bald, or that it is not the case that there is someone who is the unique king of France and bald. The first, more natural, reading, would be translated as

$$\exists x(Fx \wedge \forall y(Fy \rightarrow x = y) \wedge \neg Gx).$$

Here the existential quantifier takes scope over the negation. In the second reading, the negation scopes over the quantifier:

$$\neg\exists x(Fx \wedge \forall y(Fy \rightarrow x = y) \wedge Gx).$$

Only the second reading is the negation of 'the king of France is bald'.

In modal contexts, descriptions often give rise to *de re*/*de dicto* ambiguities. 'The Pope might have been Italian' can mean that the actual Pope, Jorge Mario Bergoglio,

might have been Italian (*de re*), but it can also mean that it might have been that some Italian person is Pope (*de dicto*). Following Russell, we would translate the first reading as

$$\exists x(Fx \wedge \forall y(Fy \rightarrow x = y) \wedge \Diamond Gx)$$

and the second as

$$\Diamond\exists x(Fx \wedge \forall y(Fy \rightarrow x = y) \wedge Gx).$$

Now return to the fact that Donald Trump might not have been identical to the 45th President of the United States. This, too, can be understood *de re* or *de dicto*. So there are two translations into $\mathfrak{L}_{MP}$. On the semantics of the previous two sections, only the *de re* translation conflicts with the fact that Trump is identical to the 45th President. The more natural *de dicto* reading does not.

Exercise 9.12

Give both translations of 'Donald Trump is the 45th President, and he might not have been the 45th President', following Russell's method. Then show that the *de re* translation is true at no world in any constant-domain model, while the *de dicto* translation is true at some worlds in some models.

I should add that not everyone has been persuaded by Russell's analysis of descriptions. For example, some hold that if there is no unique F, then 'the F is G' should count as neither true nor false. Almost everyone, however, agrees with Russell that descriptions should not be analysed as individual constants, which would not allow us to draw the *de re*/*de dicto* distinction.

So we can explain away *some* apparent counterexample to **ND′** and **WNI′** (and therefore to **ND**, **WNI**, and **NI**): those involving descriptions. But there are others.

For instance, some individuals have more than one name, and people often fail to know that these names pick out one and the same object. The ancient Babylonians wondered whether Hesperus is identical to Phosphorus. Some neighbours of Mary Ann Evans did not know that she was George Elliot. Here the English sentences involve genuine names, so $\neg\Box(a = b)$ should be an adequate translation (with an epistemic reading of the box).

Other apparent counterexamples involve the constitution of material objects. Consider a statue that is made entirely of clay. If the statue is standing on a shelf, then it's also true that a statue-shaped lump of clay is standing on the shelf, at the

exact same spot. Many philosophers would like to say that this statue-shaped lump of clay just *is* the statue. The statue and the statue-shaped lump aren't *two* material objects, mysteriously co-located in space. Rather, the statue is identical to the lump. Let's call the statue 'Goliath' and the lump 'Lumpl'. On the present account, Goliath = Lumpl, but the identity is not necessary. For the creator of the statue might have decided to add a copper hat. In that case, Lumpl would only have been a part of Goliath.

Suppose we want to allow for some such cases of contingent identity, and we see them as genuine counterexamples to **WNI′**. We then have to alter our semantics.

Exercise 9.13

Give similar (apparent) counterexamples to **ND′**.

A simple way to render **NWI′** invalid is to drop the assumption that names are always rigid. We can make names non-rigid by changing the definition of a model so that the interpretation function V assigns an individual to every name *relative to every world*.

There are good reasons to give a parallel treatment for variables. For example, if Lumpl is contingently identical to Goliath, we should be able to conclude that there is *something* that is contingently identical to Goliath: $\exists x(x = g \wedge \neg\Box(x = g))$. So variables, too, will be non-rigid, by making assignment functions world-relative.

The denotation of a singular term t is redefined as follows:

$$[t]^{M,w,g} =_{\text{def}} \begin{cases} \text{V(t,w)} & \text{if } t \text{ is a name (and } V \text{ is the interpretation function of } M) \\ \text{g(t,w)} & \text{if } t \text{ is a variable.} \end{cases}$$

In the definition of truth at a world in a model relative to an assignment, $[t]^{M,g}$ is replaced by $[t]^{M,w,g}$. Finally, we adjust the definition of an x-variant so that g' is an x-variant of g iff g' differs from g at most in the values it assigns to x *relative to any world*.

These changes can be made to either constant-domain semantics or variable-domain semantics. Either way, the result is known as **individual concept semantics**. An "individual concept" (or "intensional object") is a function from worlds to individuals. In individual concept semantics, names and variables are effectively treated as picking

out individual concepts. For example, since V assigns an individual to every name relative to every world, the full meaning of a name a in a model is given by the individual concept that maps each world w to $V(a, w)$.

Individual concept semantics has some counterintuitive features. One is that the following schema becomes valid:

$$\Box\exists x A \rightarrow \exists x \Box A$$

To see why, let's take a simple instance, such as $\Box\exists x Fx \rightarrow \exists x \Box Fx$. Suppose the antecedent is true at some world in some model. So at every accessible world v, there is at least one individual that is F (meaning that it is in $V(F, v)$). For any assignment g, we can therefore construct an x-variant g' so that $g'(x, v)$ always picks out an F-individual, for each accessible world v. Relative to g', $\Box Fx$ is true at w. So $\exists x \Box Fx$ is true at w.

This is widely regarded as problematic. In a lottery, for example, it may be certain that there is a winner: $\Box\exists x Fx$. But there may be no particular person of whom it is certain that *they* are the winner; so $\exists x \Box Fx$ should be false.

Another surprising feature of individual concept semantics is that it has no sound and complete proof procedure. There are no tree rules, or natural deduction rules, or combinations of axioms and inference rules that allow proving all and only the sentences that are true at all worlds in all models of individual concept semantics.

Fortunately, both of these problems can be avoided by putting further constraints on models. So far, we have assumed that a non-rigid name or variable may pick out Donald Trump in one world, the Eiffel tower in another, a fried egg in a third, and so on. But while one can cook up descriptions with such a gerrymandered modal profile, ordinary names and variables seem to require much more unity among their referents at different worlds. The name 'Lumpl' picks out a smallish piece of clay at every world; 'Goliath' always picks out a statue. We could therefore add to our models a restricted set of "eligible" individual concepts. Every name and every variable must be interpreted as expressing one of these concepts.

Closely related to individual concept semantics is another approach known as **counterpart semantics**. Here, too, the guiding idea is that there are different ways of re-identifying an individual at other worlds. If we think of the object on the shelf as Lumpl, and we consider a world where a copper hat has been added, we judge that Lumpl is only the clay part of the resulting statue; if we think of the same object on

the shelf as Goliath, we judge that it includes the copper hat in the counterfactual scenario. In counterpart semantics, different ways of tracking objects across worlds are represented by different *counterpart relations*. The object on the shelf at our world stands in different counterpart relations to different objects at the world where the copper hat has been added.

One way of spelling out counterpart semantics assumes that names and variables are associated with an individual and a counterpart relation. $\Box Fa$ is true (at a world in a model) iff every individual that stands in the a-relevant counterpart relation to a is F at every accessible world. Like in individual concept semantics, the set of counterpart relations eligible for quantification should be restricted.

The main advantage of counterpart semantics over individual concept semantics is that it allows for greater flexibility. For example, we don't have to assume that counterpart relations are transitive: a counterpart of a counterpart of an individual need not itself be a counterpart of the individual. This might help with the following puzzle about metaphysical possibility, due to Hugh Chandler.

My bicycle is composed of certain parts. Let's say that it could have been composed of slightly different parts. For example, it could have had a different seatpost. But it could not have been composed of *entirely* different parts. A bike composed of entirely different parts would be a different bike. Let b denote my bike and let F be a predicate that gives a detailed description of my bike as it actually is. Let F' give a detailed description of a bike that is just like mine except for the seatpost. We want to say that my bike could have fit that description. So $\Diamond F'b$. But as we make more and more changes to F, we reach a point – say F'''' – where the description could no longer have applied to my bike. So $\Diamond F''''b$ is false. But now consider what would have been the case if $F'''b$ had been true. In that case, it seems that $F''''b$ could have been true. After all, an F'''' bike differs from an F''' bike only by one part. It would be strange if the bike in a world where $F'''b$ is true could not possibly have had any different parts. So while $\Diamond F''''b$ is false at the actual world, it seems to be true at any world where $\Diamond F'''b$ is true. Since $\Diamond F'''b$ is true at the actual world, it follows that $\Diamond\Diamond F''''b$ is true as well. So we have $\Diamond\Diamond F''''b$ but not $\Diamond F''''b$.

This is puzzling, because it is widely held that metaphysical possibility is an absolute kind of possibility, with no accessibility restrictions. The logic of metaphysical possibility should then be S5. But in S5, $\Diamond\Diamond A$ entails $\Diamond A$.

In counterpart semantics, modal schemas like $\Diamond\Diamond A \rightarrow \Diamond A$ (which is equivalent to the **4** schema $\Box A \rightarrow \Box\Box A$) correspond not just to properties of the accessibility

relation but to combined properties of the accessibility relation and the counterpart relations. So one can explain what's going on in the puzzle of the bike without giving up the assumption that the accessibility relation for metaphysical modality is universal. $\Diamond\Diamond A \rightarrow \Diamond A$ is invalid because a counterpart of a counterpart of my bike need not be a counterpart of my bike.