

# Logic 2: Modal Logic

Wolfgang Schwarz

April 16, 2024

© 2024 Wolfgang Schwarz

[github.com/wo/logic2](https://github.com/wo/logic2)



This work is licensed under a [Creative Commons](https://creativecommons.org/licenses/by-nc-sa/4.0/) “Attribution-NonCommercial-ShareAlike 4.0 International” license.



# Contents

<b>Preface</b>	<b>5</b>
<b>1 Modal Operators</b>	<b>7</b>
1.1 A new language . . . . .	7
1.2 Flavours of modality . . . . .	11
1.3 The turnstile . . . . .	14
1.4 Duality . . . . .	16
1.5 A system of modal logic . . . . .	20
<b>2 Possible Worlds</b>	<b>27</b>
2.1 The possible-worlds analysis of possibility and necessity . . . . .	27
2.2 Models . . . . .	28
2.3 Basic entailment and validity . . . . .	32
2.4 Explorations in S5 . . . . .	34
2.5 Trees . . . . .	38
<b>3 Accessibility</b>	<b>47</b>
3.1 Variable modality . . . . .	47
3.2 The systems K and S5 . . . . .	51
3.3 Some other normal systems . . . . .	53
3.4 Frames . . . . .	58
3.5 More trees . . . . .	61
<b>4 Models and Proofs</b>	<b>67</b>
4.1 Soundness and completeness . . . . .	67
4.2 Soundness for trees . . . . .	69
4.3 Completeness for trees . . . . .	72
4.4 Soundness and completeness for axiomatic calculi . . . . .	76
4.5 Loose ends . . . . .	83
<b>5 Epistemic Logic</b>	<b>87</b>
5.1 Epistemic accessibility . . . . .	87
5.2 The logic of knowledge . . . . .	88

## Contents

---

5.3	Multiple Agents . . . . .	93
5.4	Knowledge, belief, and other modalities . . . . .	100
<b>6</b>	<b>Deontic Logic</b>	<b>105</b>
6.1	Permission and obligation . . . . .	105
6.2	Standard deontic logic . . . . .	107
6.3	Norms and circumstances . . . . .	111
6.4	Further challenges . . . . .	118
6.5	Neighbourhood semantics . . . . .	120
<b>7</b>	<b>Temporal Logic</b>	<b>125</b>
7.1	Reasoning about time . . . . .	125
7.2	Temporal models . . . . .	126
7.3	Logics of time . . . . .	129
7.4	Branching time . . . . .	135
7.5	Extending the language . . . . .	140
<b>8</b>	<b>Conditionals</b>	<b>145</b>
8.1	Material conditionals . . . . .	145
8.2	Strict conditionals . . . . .	148
8.3	Variably strict conditionals . . . . .	153
8.4	Restrictors . . . . .	159
<b>9</b>	<b>Towards Modal Predicate Logic</b>	<b>163</b>
9.1	Predicate logic recap . . . . .	163
9.2	Modal fragments of predicate logic . . . . .	168
9.3	Predicate logic proofs . . . . .	171
9.4	Modality de dicto and de re . . . . .	174
9.5	Identity and descriptions . . . . .	177
<b>10</b>	<b>Semantics for Modal Predicate Logic</b>	<b>183</b>
10.1	Constant domain semantics . . . . .	183
10.2	Quantification and existence . . . . .	190
10.3	Variable-domain semantics . . . . .	192
10.4	Trans-world identity . . . . .	197
<b>11</b>	<b>Answers to the Exercises</b>	<b>203</b>

# Preface

These notes are aimed at philosophy students who have taken an introductory course in formal logic. They provide an introduction to modal logic, with many philosophical applications. Along the way, they introduce general ideas that might be taught in an intermediate logic course: different methods of proof, the concept of a model, soundness and completeness, compactness, three-valued logics, free logics, supervaluation, properties of relations and orders, etc.

Chapters 1–3 introduce the standard toolkit of modal propositional logic: Kripke models, frame correspondence, some popular systems, the tableau method and axiomatic calculi. Chapter 4 goes through soundness and completeness. Chapters 5–8 turn to philosophical applications. Each of these chapters also extends the toolkit from chapter 3. Chapter 5 introduces multi-modal logics, chapter 6 ordering models and neighbourhood semantics, chapter 6 two-dimensional semantics and supervaluationism, chapter 7 conditional logics and Lewis-Stalnaker models. Chapters 9 and 10 look at some of the complexities that arise in first-order modal logic.

Apart from chapter 9, which sets the stage for chapter 10, every chapter after chapter 3 can be skipped or skimmed without affecting the accessibility of later chapters.

The best way to learn logic is by solving problems. That's why the text is frequently interrupted by exercises. As a student, you should try to do the exercises as soon as you reach them, before continuing with the text.



# 1 Modal Operators

## 1.1 A new language

Modal logic is an extension of propositional and predicate logic that is widely used 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 sometimes want to make sure that the stated conclusions really follow from the stated premises. If they do, we say that the reasoning is *valid*. By this we mean that there is no conceivable scenario in which the premises are true while the conclusions are 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 relevant logical form 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 logical form.

In natural languages like English, the logical form of sentences is not always transparent. 'Every dog barked at a tree' can mean either that there is a single tree at which

every dog barked, or that for each dog there is a tree at which it barked. The two readings have different logical consequences, so it would be good to keep them apart. Worse, 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 in which there are no ambiguities of logical form and in which all logical expressions have determinate, precise meanings. 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.) With some practice, one can also reason directly in a formal language.

Now consider the following argument.

It might be raining.

It is certain 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  $A$ .

It is certain 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 'it is certain that'. The languages of classical logic do not have such expressions.

So let's add them. Let's invent a new formal language with two new logical symbols. It doesn't matter what these look like; a popular choice is a diamond  $\Diamond$  and a box  $\Box$ . We use the diamond to formalize 'it might be that', and the box for 'it is certain that'.

If we add these symbols to the language of propositional logic, we get the standard language of modal propositional logic. If we add them to the language of predicate logic, we get the standard language of modal predicate logic. We will stick with propositional logics until chapter 9.

Let's officially define the standard language of modal propositional logic.

**Definition 1.1: The language  $\mathcal{L}_M$**

A *sentence letter* of  $\mathcal{L}_M$  is any lower-case letter of the Latin alphabet ( $a, b, c, \dots, z$ ), possibly followed by numerical subscripts ( $a_1, p_{18}, \dots$ ).

A *sentence* of  $\mathcal{L}_M$  is either a sentence letter of  $\mathcal{L}_M$  or an expression of the form  $\neg A$ ,  $(A \wedge B)$ ,  $(A \vee B)$ ,  $(A \rightarrow B)$ ,  $(A \leftrightarrow B)$ ,  $\Box A$ , or  $\Diamond A$ , where  $A$  and  $B$  are  $\mathcal{L}_M$ -sentences.

I use lower-case letters  $a, b, c, \dots$  as atomic  $\mathcal{L}_M$ -sentences and upper-case letters  $A, B, C, \dots$  when I want to talk about arbitrary  $\mathcal{L}_M$ -sentences. To reduce clutter, I generally omit outermost parentheses and quotation marks when I mention  $\mathcal{L}_M$ -symbols or sentences:  $p \wedge q$  is treated as an abbreviation of ‘ $(p \wedge q)$ ’.

**Exercise 1.1**

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

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

Having new symbols is only the beginning. We also need to lay down rules for reasoning 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  $\wedge$  that may or may not exactly match the meaning of ‘and’ in English.

The meaning of  $\wedge$  can be given by a *truth table*:

A	B	$A \wedge B$
T	T	T
T	F	F
F	T	F
F	F	F

This tells us how the truth-value of  $A \wedge B$  depends on the truth-value of  $A$  and  $B$ : the compound sentence is true iff (if and only if) both of its subsentences are true. If you know this, you know all there is to know about the meaning of  $\wedge$ . (You can see, for example, that  $A \wedge B$  does not imply anything about the temporal order of  $A$  and  $B$ .)

**Exercise 1.2**

Draw the truth tables for  $\neg$ ,  $\vee$ ,  $\rightarrow$ , and  $\leftrightarrow$ .

The sentence operators (or connectives) of classical propositional logic ( $\neg$ ,  $\wedge$ ,  $\vee$ ,  $\rightarrow$ , and  $\leftrightarrow$ ) are all truth-functional. Recall that an operator is **truth-functional** if the truth-value of a compound sentence formed by applying the operator to other sentences is always determined by the truth-value of these other sentences. The truth tables for the classical operators spell out this dependence. They tell us how to compute the truth-value of a compound sentence from the truth-values of its constituents.

The diamond operator can't be truth-functional if it is supposed to mean anything like 'it might be that' in English. To see why, note first that 'it might be that  $P$ ' can be true if  $P$  is true, but also if  $P$  is false. 'It might be raining' doesn't entail that it is actually raining, nor that it isn't raining. It merely says that our evidence is compatible with rain. Now, if the diamond were truth-functional, then what would follow from the fact that  $\Diamond p$  is *sometimes* true when  $p$  is true? It would follow that  $\Diamond p$  is *always* true when  $p$  is true. (Make sure you understand why.) Likewise, from the fact that  $\Diamond p$  is sometimes true when  $p$  is false, it would follow that  $\Diamond p$  is true whenever  $p$  is false.  $\Diamond p$  would be a logical truth. But 'it might be raining' is surely not a logical truth.

If an operator isn't truth-functional, its meaning can't be defined by a truth table. The standard approach to defining 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.

**Exercise 1.3**

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 ...

## 1.2 Flavours of modality

‘It might be that’ and ‘it is certain that’ express an *epistemic* kind of possibility and necessity, related to evidence and knowledge. There are other kinds – or *flavours* – of possibility and necessity.

Consider ‘John must leave’. This expresses a kind of necessity, but it would typically not be understood as a statement about the available evidence. On its most natural interpretation, it says that some relevant norms require John to leave. This flavour of necessity is called *deontic* (from Greek *deontos*: ‘of that which is binding’).

Other statements about possibility and necessity are neither deontic nor epistemic. If I say that you can’t travel from Auckland to Sydney by train, I don’t just mean that my information implies that you won’t make that journey; nor do I mean that you’re not permitted to make it. Rather, I mean that relevant circumstances in the world – such as the presence of an ocean between Auckland and Sydney – preclude the journey. This flavour of modality is sometimes called *circumstantial*. It comes in many sub-flavours, depending on what kinds of circumstances are in play.

Each of these flavours of modality corresponds to a branch of modal logic. *Epistemic logic* formalizes reasoning about knowledge and information. *Deontic logic* deals with norms, permissions, and obligations. 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 this label is also not used widely, and it is used for different things by different authors.

Confusingly, some philosophers use ‘modal logic’ for the logic of a certain sub-flavour of circumstantial modality, known as *metaphysical* modality. Metaphysical modality is concerned with what is or isn’t compatible with the nature of things. We will follow the more common practice of using ‘modal logic’ as an umbrella term that covers all the applications I have mentioned, as well as many others.

We will take a closer look at epistemic logic in chapter 5 and at deontic logic in chapter 6. In chapter 7 we are going to study a branch of modal logic called *temporal logic* that is concerned with reasoning about time. Chapter 8 is on *conditional logic*. Here we will introduce (non-truth-functional) two-place operators that are meant to formalise certain ‘if ...then ...’ constructions in English. In chapter 4, we will briefly look at *provability logic*, which investigates formal properties of mathematical provability. What unifies the different branches of modal logic is not a particular subject matter, but a loosely defined collection of abstract ideas and techniques that turn out to be useful in all these applications.

When we study some flavour of possibility or necessity, the diamond  $\Diamond$  is generally

used for the relevant kind of possibility and the box  $\Box$  for the corresponding kind of necessity. In this context, you may pronounce the diamond ‘it is possible that’ and the box ‘it is necessary that’. In general, however, I would recommend pronouncing the diamond ‘diamond’ and the box ‘box’.

Different interpretations of the box and the diamond often motivate different rules for reasoning with these expressions. Consider, for example, the inference from  $\Box p$  to  $p$ . If the box expresses a circumstantial kind of necessity, then this inference is plausibly valid: if the circumstances ensure that something is the case, then it really is the case. On a deontic reading of the box, by contrast, the inference is invalid. We can easily imagine scenarios in which, say, it is required that all library books are returned on time ( $\Box p$ ) and yet it is not the case that all library books are returned on time ( $\neg p$ ).

So we can’t say, once and for all, whether  $\Box p$  entails  $p$ . We will develop different “logics” or “systems” of modal logic. In some systems, the inference is valid, in others it is invalid.

The diamond and the box are sentence operators. English expressions for necessity and possibility often don’t have this form. We can talk about what’s necessary or possible using ‘must’, ‘might’, or ‘can’, which are (auxiliary) verbs. We can also use adjectives like ‘feasible’, ‘certain’, and ‘obligatory’, or adverbs like ‘possibly’, ‘certainly’, and ‘inevitably’.

When translating from English into  $\mathcal{L}_M$ , it is often helpful to first paraphrase the English sentence with ‘it is necessary that’ and ‘it is possible that’ (or other suitable sentence operators). For example,

You can’t go from Auckland to Sydney by train

might be paraphrased as

It is not possible [in light of relevant circumstances] that you go from Auckland to Sydney by train

An adequate translation is  $\neg\Diamond p$ , where  $p$  represents ‘you go from Auckland to Sydney by train’ and the diamond represents the relevant kind of circumstantial possibility.

#### Exercise 1.4

Translate the following sentences, as well as possible, into  $\mathcal{L}_M$ , 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.
- (e) If the murderer escaped through the window, there might be traces on the ground.

#### Exercise 1.5

Translate the following sentences, as well as possible, into  $L_M$ , 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.

#### Exercise 1.6

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

- (a) I could have studied architecture.
- (b) The bridge is fragile.
- (c) I can't hear you if you're talking to me from the kitchen.
- (d) If you have a smartphone, you can use an electronic ticket.

Special care is required when translating English sentences that contain both modal expressions and an 'if' clause. The surface form of English can be misleading. A good strategy is to first rephrase the English sentence so that it no longer contains any conditional expression, then translate that paraphrase. The paraphrase, and therefore the translation, will often sound rather unlike the original sentence, but that's OK. What's important is that it has the same truth-conditions. There should be no conceivable scenario in which the original sentence is true and the paraphrase (or translation) false, or the other way round.

### 1.3 The turnstile

In section 1.1, I said 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, I said, if it is valid “in virtue of its logical form”. Can we make this more precise?

Consider this English argument.

Some cats are black.

Therefore: Some animals are black.

The argument is valid, but not logically valid. Its validity turns on the meaning of ‘cat’, which we don’t consider a logical expression.

To bring out how the argument’s validity depends on the meaning of ‘cat’, we can imagine a language that is much like English except that ‘cat’ means *chair*. In this language, the argument just displayed is invalid. It is invalid because there are conceivable scenarios in which there are black chairs but no black animals. In any such scenario, the argument’s premise is true (in our imaginary language) while the conclusion is false.

When we say that an argument is valid “in virtue of its logical form”, we mean that its validity does not depend on the meaning of the non-logical expressions. In other words, there is no conceivable scenario in which the premises are true and the conclusion is false, *no matter what meaning we assign to the non-logical expressions*.

The concept of validity for arguments is closely related to that of entailment. If an argument is valid, we say that the premises entail the conclusion. If an argument is logically valid, we say that the premises logically entail the conclusion. In logic, we’re interested in logical entailment. We adopt the following definition.

#### Definition 1.2

Some sentences  $\Gamma$  (‘gamma’) **(logically) entail** a sentence  $A$  iff there is no conceivable scenario in which all sentences in  $\Gamma$  are true and  $A$  is false, under any interpretation of the non-logical expressions.

Instead of saying that the sentences  $\Gamma$  logically entail  $A$ , we also say that  $A$  is a *logical consequence* of  $\Gamma$ , or that  $A$  *logically follows from*  $\Gamma$ . Two sentences are (*logically*) *equivalent* if either logically follows from the other.

Logicians often use the symbol ‘ $\models$ ’ (the “double-barred turnstile”) for entailment. The claim that  $\Box(p \rightarrow q)$  and  $\Box p$  together entail  $q$ , for example, could be expressed as

$$\Box(p \rightarrow q), \Box p \models q.$$

This is not a sentence of  $\mathcal{L}_M$ . The comma and the turnstile belong to the **meta-language** we use to talk about the **object language**  $\mathcal{L}_M$ . (The rest of our meta-language is mostly English.) We use the turnstile to express a certain relationship between  $\mathcal{L}_M$ -sentences, not to construct further  $\mathcal{L}_M$ -sentences.

### Exercise 1.7

What do you think of this simpler alternative to definition 1.2? “Sentences  $\Gamma$  entail a sentence  $A$  iff there is no interpretation of non-logical expressions that renders all sentences in  $\Gamma$  true and  $A$  false.”

The following fact about logical consequence often proves useful.

**Observation 1.1:** If  $A$  and  $B$  are sentences and  $\Gamma$  is a (possibly empty) list of sentences, then

$$\Gamma, A \models B \text{ iff } \Gamma \models A \rightarrow B.$$

*Proof.* Look at the statement on the right-hand side of the ‘iff’. ‘ $\Gamma \models A \rightarrow B$ ’ says that there is no conceivable scenario in which all sentences in  $\Gamma$  are true while  $A \rightarrow B$  is false, under any interpretation of the non-logical expressions. By the truth-table for ‘ $\rightarrow$ ’,  $A \rightarrow B$  is false iff  $A$  is true and  $B$  is false. So we can rephrase the statement on the right-hand side as saying that there is no conceivable scenario and interpretation that makes all sentences in  $\Gamma$  true and  $A$  true and  $B$  false. That’s just what the statement on the left-hand side asserts.  $\square$

Observation 1.1 tells us that if we start with a claim of the form  $A_1, A_2, A_3 \dots \models B$ , we can always generate an equivalent claim by moving the turnstile to the left of the sentence that precedes it and putting an arrow in its original place. For example, instead of

$$\Box(p \rightarrow q), \Box p \models \Box q$$

we can equivalently say

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

We can go further to

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

This says that  $\Box(p \rightarrow q) \rightarrow (\Box p \rightarrow \Box q)$  logically follows from no premises at all. A sentence that follows from no premises is 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.)

Sentence validity is implicitly covered by definition 1.2, using an empty list of sentences for  $\Gamma$ . But it's worth making the definition more explicit.

### Definition 1.3

A sentence  $A$  is **valid** (for short,  $\models A$ ) iff there is no conceivable scenario in which  $A$  is false, under any interpretation of the non-logical expressions.

Make sure you don't confuse the arrow with the turnstile. It's not just that the two symbols belong to different languages – one to  $\mathcal{L}_M$ , the other to our meta-language. They also have very different meanings.  $p \rightarrow q$  is true iff either  $p$  is false or  $q$  is true (or both).  $p \models q$ , on the other hand, is true iff there is no conceivable scenario in which  $p$  is true and  $q$  is false, under any interpretation of  $p$  and  $q$ . Nonetheless, there is an important connection between the arrow and the turnstile:  $A \models B$  is *true* iff  $A \rightarrow B$  is *valid*.

The definitions of this section are still somewhat imprecise. Eventually we will want to prove various claims about entailment and validity. To this end, we will need to give rigorous meanings to 'conceivable scenario' and 'interpretation of non-logical expressions'. Let's leave this task until the next chapter.

## 1.4 Duality

'Neville can't be the murderer', says Watson. His claim could be paraphrased as 'it is not possible that Neville is the murderer'. This suggests that  $\neg\Diamond p$  is an adequate translation (where  $p$  expresses that Neville is the murderer). But Watson's claim might also be paraphrased as 'it is certain that Neville is not the murderer', which we might translate as  $\Box\neg p$ .

The two paraphrases are plausibly equivalent. In general, 'it is not (epistemically) possible that  $A$ ' seems to say the same as 'it is certain that not  $A$ '. Similarly, 'it is not certain that  $A$ ' arguably says the same as 'it is possible that not  $A$ '.

Whether or not the equivalence holds in English, we stipulate that it holds in  $\mathcal{L}_M$ : for any  $\mathcal{L}_M$ -sentence  $A$ ,

(Dual1)  $\neg\Diamond A$  is equivalent to  $\Box\neg A$ ;

(Dual2)  $\neg\Box A$  is equivalent to  $\Diamond\neg A$ .

Operators that stand in the relationship expressed by (Dual1) and (Dual2) are called **duals** of each other. There is a convention in modal logic to use the symbols  $\Box$  and  $\Diamond$  only for concepts that are duals of each other.

### Exercise 1.8

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 at some point the case that ...
- (f) It will at some point 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 ...

(Dual1) implies that  $\neg\Diamond\neg p$  is equivalent to  $\Box\neg\neg p$ , choosing  $\neg p$  as the sentence  $A$ . In standard modal logic, logically equivalent expressions are interchangeable. So we can simplify  $\Box\neg\neg p$  to  $\Box p$ , drawing on the equivalence between  $\neg\neg p$  and  $p$ . So  $\neg\Diamond\neg p$  is equivalent to  $\Box p$ .

The same reasoning could be applied to any other sentence  $A$  in place of  $p$ . (Dual1) therefore implies that for any sentence  $A$ ,

$\Box A$  is equivalent to  $\neg\Diamond\neg A$ .

In the same way, (Dual2) implies that (for any sentence  $A$ )

$$\Diamond A \text{ is equivalent to } \neg \Box \neg A.$$

This shows that the box and the diamond can be defined in terms of one another. We could have used a language whose only primitive modal operator is the box, and read  $\Diamond A$  as an abbreviation of  $\neg \Box \neg A$ . Alternatively, we could have used the diamond as the only primitive modal operator and read  $\Box A$  as an abbreviation of  $\neg \Diamond \neg A$ .

### Exercise 1.9

Which of these sentences are equivalent to  $\Diamond \Diamond \neg p$ ? (a)  $\Diamond \neg \Diamond p$ , (b)  $\Diamond \neg \Box p$ , (c)  $\neg \Box \Diamond p$ , (d)  $\neg \Diamond \Box p$ , (e)  $\neg \Box \Box p$

A digression: you might think that there is another connection between ‘possible’ and ‘necessary’. When we say that something is possible (or that it might be the case), we often convey that it is not necessary (or not certain). This suggests that  $\Diamond p$  entails  $\neg \Box p$ . We’ve just assumed, however, that  $\Diamond p$  is equivalent to  $\neg \Box \neg p$ . If  $\Diamond p$  entails  $\neg \Box p$ , we would have to conclude that  $\neg \Box \neg p$  entails  $\neg \Box p$ . By contraposition, we could infer that  $\Box p$  entails  $\Box \neg p$ . But ‘it is necessary that  $P$ ’ surely doesn’t entail ‘it is necessary that not- $P$ ’!

We have to reject either the duality of ‘possible’ and ‘necessary’ or the apparent entailment from ‘possible’ to ‘not necessary’. On reflection, the case for duality is stronger. There is a good explanation of why ‘possible’ often *appears* to entail ‘not necessary’ even if it actually doesn’t.

Take an example. Suppose Watson says ‘Neville might be the murderer’. Let’s assume that ‘might’ is the dual of ‘certain’, so that ‘it might be that  $P$ ’ is equivalent to ‘it is not certain that not  $P$ ’. On this interpretation, what Watson said – that Neville might be the murderer – is merely that it isn’t certain that Neville is *not* the murderer. It may well be certain that Neville *is* the murderer. Why, then, does his statement convey that Neville’s guilt is an open question?

Well, suppose Watson had known that Neville is the murderer. In that case, he shouldn’t have said ‘Neville might be the murderer’. These words would still have been true – or so we assume – but they would not have been helpful. Watson would have been in a position to say something more informative: that Neville is the murderer, or that he is known to be the murderer. We generally assume that speakers are trying to be helpful, that they are not hiding relevant information. Assuming that Watson is trying to be helpful, his *statement* that Neville might be the murderer implies that he considers Neville’s guilt

an open question. This follows not from *what he said*, but from the fact *that he said it*, together with the assumption that he is trying to be helpful.

This kind of effect is studied in the field of pragmatics, where it is known as a *scalar implicature*. Scalar implicatures arise when an utterance of a logically weaker sentence conveys that a certain stronger sentence is false. ‘Some students passed the test’, for example, conveys that not all students passed the test, although the statement would be true even if all students had passed. In that case, however, it would not have been helpful: the speaker should have used ‘all students passed’. End of digression.

I want to say a little more about duality. To do so, I need to introduce the concept of a schema.

Formally, a **schema** (for  $\mathcal{L}_M$ -sentences) is simply an  $\mathcal{L}_M$ -sentence with upper-case schematic variables in place of sentence letters. Every  $\mathcal{L}_M$ -sentence that results from a schema by (uniformly) replacing the schematic variables with object-language sentences is called an **instance** of the schema.

$\Box A \rightarrow A$ , for example, is a schema. Three of its instances are  $\Box p \rightarrow p$  and  $\Box(p \vee q) \rightarrow (p \vee q)$  and  $\Box\Box p \rightarrow \Box p$ . The sentence  $\Box p \rightarrow q$  is not an instance: the same schematic variable must always be replaced by the same object-language sentence. (That’s what I meant by “uniformly”.)

### Exercise 1.10

Which of the following expressions are instances of  $\Box(A \rightarrow \Diamond(A \wedge B))$ ?

- (a)  $\Box(p \rightarrow \Diamond(q \wedge p))$
- (b)  $\Box(\Diamond p \rightarrow \Diamond(\Diamond p \wedge p))$
- (c)  $\Box\Box(p \rightarrow \Diamond(p \wedge q))$
- (d)  $\Box((p \rightarrow \Diamond(p \wedge q)) \rightarrow \Diamond((p \rightarrow \Diamond(p \wedge q)) \wedge \Diamond p))$
- (e)  $\Box((A \wedge C) \rightarrow \Diamond((A \wedge C) \wedge (B \wedge C)))$

Schemas are useful when we want to talk about all  $\mathcal{L}_M$ -sentences of a certain form. In the next section, for example, we are going to define a system of modal logic by giving a list of schemas all instances of which are considered valid.

Now compare the schemas  $\Box A \rightarrow A$  and  $A \rightarrow \Diamond A$ . Given the duality of the box and the diamond, and the fact that logically equivalent expressions can be freely exchanged for one another, we can show that *every instance of one of them is equivalent to an instance of the other*. In this sense, the two schemas are equivalent. And because their equivalence relies on the duality of the box and the diamond, the two schemas are called *duals* of one another.

To see why every instance of  $\Box A \rightarrow A$  is equivalent to an instance of  $A \rightarrow \Diamond A$ , take a simple instance:  $\Box p \rightarrow p$ . By the truth-table for the arrow, this is equivalent to  $\neg p \rightarrow \neg \Box p$ . By (Dual2),  $\neg \Box p$  is equivalent to  $\Diamond \neg p$ . So  $\neg p \rightarrow \neg \Box p$  is equivalent to  $\neg p \rightarrow \Diamond \neg p$ . And this is an instance of  $A \rightarrow \Diamond A$ . The same line of reasoning obviously works for any other sentence in place of  $p$ , and a similar line of reasoning shows the converse, that every instance of  $A \rightarrow \Diamond A$  is equivalent to an instance of  $\Box A \rightarrow A$ .

It's crucial that we're talking about schemas here. We have not shown that the *sentence*  $\Box p \rightarrow p$  is equivalent to  $p \rightarrow \Diamond p$ . In fact, the duality principles and the replacement of equivalents don't suffice to show that these sentences are equivalent.

The equivalence of the *schemas*, however, is enough to show that it doesn't matter which of them we use when we list schemas to define a logic. We can say that all instances of  $\Box A \rightarrow A$  are valid in a certain logic, or we can say that all instances of  $A \rightarrow \Diamond A$  are valid – it amounts to the same thing, because every instance of either schema is equivalent to an instance of the other.

The equivalence between  $\Box A \rightarrow A$  and  $A \rightarrow \Diamond A$  is an example of a more general pattern. Any schema with an arrow ( $\rightarrow$  or  $\leftrightarrow$ ) as the only truth-functional operator can be converted into an equivalent schema – its **dual** – by swapping antecedent and consequent and replacing every box with a diamond and every diamond with a box.

### Exercise 1.11

Find the duals of (a)  $\Box A \rightarrow \Box \Box A$ , (b)  $\Diamond A \rightarrow \Box \Diamond A$ , (c)  $\Box A \rightarrow \Diamond A$ .

### Exercise 1.12

A proposition is *contingent* if it is neither necessary nor impossible. Let  $\nabla$  be a sentence operator for 'it is contingent that'. Reading the box as 'it is necessary that' and the diamond as 'it is possible that', try to find

- (a) a sentence whose only modal operator is  $\Box$  that is equivalent to  $\nabla p$ ;
- (b) a sentence whose only modal operator is  $\Diamond$  that is equivalent to  $\nabla p$ ;
- (c) a sentence whose only modal operator is  $\nabla$  that is equivalent to  $\Box p$ .

## 1.5 A system of modal logic

Whether a sentence is logically valid, or logically entailed by other sentences, never depends on the meaning of the non-logical expressions. But it may well depend on the meaning of the logical expressions. In modal logic, the box and the diamond are

treated as logical expressions, but their interpretation varies from application to application. Sometimes the box means epistemic necessity, sometimes it means deontic necessity, sometimes it means something else. As I mentioned in section 1.2, this has the consequence that we need to distinguish different “systems of modal logic”. In some applications, we want  $\Box p$  to entail  $p$ , in others we don’t.

Suppose, now, that we want to fully spell out one of these “systems”. We want to completely specify which  $\mathcal{L}_M$ -sentences are valid, and which are entailed by which others, on a particular understanding of the modal operators.

There are many ways of approaching this task. We could, for example, define precise notions of conceivable scenarios and interpretations and apply the definitions of the previous section. But let’s choose a more direct route. When we think about circumstantial necessity, we can intuitively see that  $\Box p$  entails  $p$ , without going through sophisticated considerations about scenarios and interpretations. Assume, then, that we simply start with direct judgements about entailment and validity.

We still face a problem. There are infinitely many  $\mathcal{L}_M$ -sentences. We can’t look at every sentence and argument one by one. We need to find some shortcuts.

We can begin by drawing on a consequence of observation 1.1. Above I said that in order to spell out a system of modal logic, we need to specify (i) which  $\mathcal{L}_M$ -sentences are valid and (ii) which  $\mathcal{L}_M$ -sentences are entailed by which others. Observation 1.1 tells us that we can ignore part (ii) of the task. Once we have fixed which sentences are valid, we have implicitly also fixed which sentences entail which others. If, for example, we decide that  $\Box p \rightarrow p$  is valid, we have also decided that  $\Box p$  entails  $p$ .

Our task of spelling out a system of modal logic therefore reduces to the task of specifying which  $\mathcal{L}_M$ -sentences are valid. That’s why a **system of modal logic** is usually defined simply as a set of  $\mathcal{L}_M$ -sentences.

To make this more concrete, let’s look at a particular sub-flavour of circumstantial necessity, sometimes called *historical necessity*. Something is historically necessary if it is “settled”: it is true and there is nothing anyone can do about it. Facts about the past are plausibly settled. Nothing we can do is going to make a difference to what happened yesterday. By contrast, some facts about the future are intuitively “open”.

Let’s use the box to formalise this (admittedly vague) concept of historical necessity. So  $\Box p$  says that  $p$  is settled. Since the diamond is the dual of the box,  $\Diamond p$  expresses that it is not settled that  $p$  is false. In other words,  $p$  is either open or settled as true.

Our task is to specify all  $\mathcal{L}_M$ -sentences that are valid on this understanding of the box and the diamond. This will give us a system of modal logic, a set of  $\mathcal{L}_M$ -sentences that are valid on a certain interpretation of the box and the diamond. We want to know which sentences are in the system – for short, which sentences are “in” – and which are not.

If the box expresses historical necessity then  $\Box p$  clearly entails  $p$ . So  $\Box p \rightarrow p$  is in. There is nothing special here about the sentence  $p$ . Whatever is settled is true. Every instance of the schema  $\Box A \rightarrow A$  is in. (As mentioned in section 1.4, it follows that every instance of  $A \rightarrow \Diamond A$  is in as well.)

In the same vein, we may now look at other schemas. Arguably, all instances of the following schemas – listed here with their conventional names – are valid, and therefore in our target system:

- (Dual)  $\neg \Diamond A \leftrightarrow \Box \neg A$
- (T)  $\Box A \rightarrow A$
- (K)  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
- (4)  $\Box A \rightarrow \Box \Box A$
- (5)  $\Diamond A \rightarrow \Box \Diamond A$

(Dual) corresponds to the duality principle (Dual1) from section 1.4. Its instances are guaranteed to be valid by the fact that we have introduced the diamond as the dual of the box.

We've already talked about (T).

(K) is a little easier to understand as a claim about entailment:

$$\Box(A \rightarrow B), \Box A \models \Box B.$$

On our present interpretation, this says that if a material conditional  $A \rightarrow B$  is settled, and its antecedent  $A$  is settled, then its consequent  $B$  is guaranteed to be settled as well. Why should we accept this? Let  $A$  and  $B$  be arbitrary propositions, and assume that  $A \rightarrow B$  and  $A$  are both settled. It follows that they are both true. Since  $A \rightarrow B$  and  $A$  entail  $B$ , it follows that  $B$  is true as well. Could it be that  $B$  is true but open? Arguably not: If we could bring about a situation in which  $B$  is false then we could also bring about a situation in which either  $A \rightarrow B$  or  $A$  is false, since one of these is guaranteed to be false in any situation in which  $B$  is false. The assumption that  $A \rightarrow B$  and  $A$  are settled therefore implies that  $B$  is settled. So all instances of (K) are in.

(4) and (5) assert that facts about what is settled are themselves settled. (4) says that if something is settled then it is settled that it is settled. (5) says that if something is not settled then it is settled that it is not settled. Here it is important that we adopt a consistent point of view. It is easy to think of situations in which something is open to us (say, we could read a certain letter) and we can do something (say, burn the letter) that would make it no longer open. This doesn't contradict (5), since (5) concerns what is

open and settled *now*. If something is now open, then arguably there is nothing we can do that would change the fact that it is now open. Likewise, if something is now settled, then arguably there is nothing we can do that would change the fact that it is now settled.

I could have listed further schemas. For example, whenever a conjunction is settled, then both its conjuncts are plausibly settled as well. So every instance of  $\Box(A \wedge B) \rightarrow (\Box A \wedge \Box B)$  should be in. There are, in fact, infinitely many further schemas, not covered by the five above, whose instances belong to our target system.

That's the bad news. The good news is that we don't need to list any of them. We can replace the whole lot by specifying two rules for generating new sentences from sentences we have already classified as "in".

The first of these rules captures the plausible thought that anything that follows from a valid sentence by classical (non-modal) propositional logic is itself valid. Since we've decided that  $\Box p \rightarrow p$  is valid (in the logic of historical necessity), we can, for example, infer that  $(\Box p \rightarrow p) \vee q$  is also valid, because  $A \vee B$  follows from  $A$  in classical propositional logic. Our system of modal logic thereby becomes an **extension** of classical propositional logic.

To state the rule concisely, let  $\Gamma \models_0 A$  mean that  $A$  follows from  $\Gamma$  in classical propositional logic – as can be determined, for example, by the truth table method. Then our rule says that for any list of sentences  $\Gamma$  and any sentence  $A$ ,

(CPL) If  $\Gamma \models_0 A$  and all members of  $\Gamma$  are in, then  $A$  is in.

As a special case, (CPL) implies that every propositional tautology is "in", since tautologies follow in classical propositional logic from any premises whatsoever (and even from no premises).

Our second rule reflects the idea that all logical truths are settled: For any sentence  $A$ ,

(Nec) If  $A$  is in, then  $\Box A$  is in.

And now we're done. I claim – and this may seem rather mysterious at the moment – that there is a natural understanding of historical necessity (of 'settled') on which the sentences that are valid in the logic of historical necessity are precisely the sentences that can be generated from instances of (T), (K), (4), (5) and (Dual) by (CPL) and (Nec). (In fact, (4) is redundant: any instance of (4) can be derived from the remaining axioms and rules.)

The system of modal logic defined by these schemas and rules is perhaps the best known of all systems of modal logic. Its conventional name is 'S5' because it was in-

roduced as the fifth system in an influential list of systems published by C.I. Lewis and C.H. Langford in 1932.

Other systems of modal logic can be defined by different schemas or rules. Lewis and Langford’s system S4, for example, is defined by (T), (K), (4), (Dual), (CPL) and (Nec), without (5). This system is adequate for other interpretations of the box and the diamond, where we don’t want to treat all instances of (5) as valid.

### Exercise 1.13

Instead of reading the box as ‘it is settled that’, we might give it one of these interpretations (with the diamond defined as the box’s dual):

- (a) it is true that
- (b) it is false that
- (c) it is either true or false that
- (d) it is logically true that

For each of these interpretations, evaluate whether the schemas (T), (K), (4), (5), and the rules (CPL) and (Nec) are plausible.

Remember that a system of modal logic is just a set of  $\mathcal{L}_M$ -sentences. I have defined the system S5 in terms of (T), (K), (4), (5), or (Dual), (CPL) and (Nec), but the same system can be defined by many other combinations of schemas and rules. (Lewis and Langford used a very different definition.)

The schemas and rules that I have chosen are called an **axiomatisation** of S5. The schemas – or more precisely, their instances – are called **axioms** because they are the starting points if we want to show that a sentence is in the system.

To illustrate this point, think of how we could show that  $\Box(p \wedge q) \rightarrow \Box p$  is in S5 (that it is “S5-valid”). The sentence is not an instance of any of the schemas I have listed. Instead, we may start with the non-modal sentence  $(p \wedge q) \rightarrow p$ . This is a propositional tautology, so (CPL) tells us that it is in S5. By (Nec), it follows that  $\Box((p \wedge q) \rightarrow p)$  is in S5 as well. Since all instance of (K) are in S5, the system contains

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

By Modus Ponens,  $\Box((p \wedge q) \rightarrow p)$  and  $\Box((p \wedge q) \rightarrow p) \rightarrow (\Box(p \wedge q) \rightarrow \Box p)$  entail our target sentence  $\Box(p \wedge q) \rightarrow \Box p$ . By (CPL), this means the target sentence is also in S5.

Here is a more streamlined presentation of this line of reasoning.

1.  $(p \wedge q) \rightarrow p$  (CPL)
2.  $\Box((p \wedge q) \rightarrow p)$  (1, Nec)
3.  $\Box((p \wedge q) \rightarrow p) \rightarrow (\Box(p \wedge q) \rightarrow \Box p)$  (K)
4.  $\Box(p \wedge q) \rightarrow \Box p$  (2, 3, CPL)

We can use the same streamlined format to show that, say,  $\Box p \rightarrow \Diamond p$  is S5-valid.

1.  $\Box \neg p \rightarrow \neg p$  (T)
2.  $\neg \Diamond p \leftrightarrow \Box \neg p$  (Dual)
3.  $\neg \Diamond p \rightarrow \neg p$  (1, 2, CPL)
4.  $p \rightarrow \Diamond p$  (3, CPL)
5.  $\Box p \rightarrow p$  (T)
6.  $\Box p \rightarrow \Diamond p$  (4, 5, CPL)

These annotated lists look a lot like proofs. They *are* proofs. Every axiomatisation of a logical system defines a corresponding **axiomatic calculus**. A proof in an axiomatic calculus is simply a list of sentences each of which is either an axiom or follows from earlier sentences in the list by one of the rules. (The annotations on the right are not officially part of the proof. They are added to help understand where the lines come from.)

#### Exercise 1.14

Try to find axiomatic proofs showing that the following sentences are in S5.

- (a)  $\Box(\Box p \rightarrow p)$
- (b)  $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$
- (c)  $\Diamond \neg p \leftrightarrow \neg \Box p$

#### Exercise 1.15

In the axiomatic calculus for S5, (Nec) allows us to derive  $\Box A$  from  $A$ . Someone might object that this inference is obviously invalid, since a sentence might be true without being necessarily true. Can you explain why (Nec) is an acceptable rule in the axiomatic calculus for S5?

The axiomatic method is the oldest formal method of proof. It has many virtues, but user-friendliness is not among them. Even simple facts are often hard to prove in an axiomatic calculus. In the next chapter, we will meet a different method that is much easier to use.

## 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 1940s-60s. The core idea of possible-worlds semantics is to analyze modal notions in terms of truth at possible worlds. In its simplest form, the analysis goes like this:

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 jargon, a **possible world** is a maximally specific possibility. An example of a possible world is the **actual world** – the totality of everything that is the case. In the actual world, light travels faster than sound and the Conservatives are in government. In other possible worlds, sound travels faster than light and Labour is in government.

The possible-worlds analysis translates modal statements into quantificational statements about possible worlds. You may feel uneasy about this. Talking about merely possible worlds may strike you as fanciful and unscientific. Besides, you may wonder if anything is really gained by the translation, since we now face the question what sorts of worlds should be classified as “possible”.

Remember that there are different flavours of modality. A proposition might be epistemically possible, historically possible, metaphysically possible, and so on. If we want to analyse all these kinds of possibility in terms of possible worlds, we need different flavours of worlds. There must be epistemically possible worlds, historically possible worlds, metaphysically possible worlds, etc. And if we ask how these types of worlds are defined it looks like we have to turn back to relevant features of the actual world. The ultimate reason why you can’t go from Auckland to Sydney by train is surely that there is no suitable train line here in our world, not that you don’t make the journey in some non-actual worlds.

These objections cast doubt on the possible-worlds analysis as a piece of reductive metaphysics. But the metaphysics of modality is not our topic. When we use the possible-world analysis, we don't assume that the translation in terms of possible worlds reveals the metaphysical grounds of the original modal statements. We merely assume that the original statements can be paraphrased in the fanciful language of possible worlds.

For a first glimpse of why this might be useful, consider the following hypothesis.

$$\Box\Diamond\Box p \models \Box p$$

Is this true? If something is necessarily possibly necessary, does it follow that it is necessary? Hard to say. We know that  $A$  logically entails  $B$  iff there is no conceivable scenario in which  $A$  is true and  $B$  false, under any interpretation of the non-logical expressions. The problem is that it is not obvious what a scenario would have to look like for  $\Box\Diamond\Box p$  to be true, under a given interpretation of  $p$ .

The possible-worlds analysis can help clear things up. By the possible-worlds analysis,  $\Box\Diamond\Box p$  says that  $\Diamond\Box p$  is true at every possible world. The hypothesis that  $\Box\Diamond\Box p$  is true in a scenario therefore reduces to the hypothesis that  $\Diamond\Box p$  is true at every world in the scenario.  $\Diamond\Box p$  says that  $\Box p$  is true at some world. So if  $\Diamond\Box p$  is true at every world in a scenario then  $\Box p$  is true at some world in the scenario. And if  $\Box p$  is true at some world in a scenario then  $p$  is true at every world in the scenario. This is just what  $\Box p$  says. So whenever  $\Box\Diamond\Box p$  is true in a scenario (under some interpretation of  $p$ ), then  $\Box p$  is true in that scenario (under that interpretation). We've shown that  $\Box\Diamond\Box p$  entails  $\Box p$ .

### Exercise 2.1

Explain, in the same informal manner, why  $\Diamond p$  does not entail  $\Box p$ , assuming the possible-worlds analysis of the box and the diamond.

## 2.2 Models

In section 1.3, I defined validity and entailment in terms of scenarios and interpretations. A sentence is valid, I said, iff it is true in every conceivable scenario under every interpretation of the non-logical expressions. This is a little vague. What, exactly, is a conceivable scenario, and what counts as a relevant interpretation? Also, scenarios and interpretations are unwieldy objects. It is difficult to give a full description of a scenario and an interpretation. Fortunately, most of the details are irrelevant if all we care about is which  $\mathcal{L}_M$ -sentences are true and which are false in a scenario under a particular in-

interpretation. This observation will lead us to a more precise definition of validity and entailment.

Suppose I tell you the following about a scenario  $S$  and an interpretation  $I$  of the sentence letters.

There are three worlds in  $S$ ,  $w_1$ ,  $w_2$ , and  $w_3$ . Under the interpretation  $I$ , the sentence  $p$  expresses a proposition that is true at  $w_1$ , false at  $w_2$ , and true at  $w_3$ . All other sentence letters express propositions that are false at all three worlds.

This tells you almost nothing about what the scenario looks like. You don't know if  $w_1$  is a world at which it is currently raining. You don't know who is in government at  $w_2$ . You also don't know what the sentence letters mean under my interpretation. Does  $p$  mean that it is raining? That Labour is in government? I haven't told you. Yet the sparse information I have given is enough to determine the truth-value of every  $\mathcal{L}_M$ -sentence at every world.

### Exercise 2.2

Which of the following sentences are true at  $w_1$  in my scenario  $S$  under my interpretation  $I$ ?

- (a)  $\neg p$
- (b)  $\neg p \rightarrow \Box p$
- (c)  $\Box p$
- (d)  $\Diamond \Box p$
- (e)  $\Diamond \Diamond p \vee \Diamond \Box p$
- (f)  $\Box(\Box p \rightarrow p)$

A joint representation of a scenario and an interpretation (of non-logical expressions) that contains just enough information to determine the truth-value of every sentence is called a **model**. Just as a model airplane often leaves out important aspects of a real airplane – the motor, the seats, etc. – models in logic leave out many important aspects of the scenarios and interpretations they represent.

Adopting the simple possible-worlds analysis of the box and the diamond, we can define a model for  $\mathcal{L}_M$  as consisting of two parts. First, a model must specify a set of things we call “worlds”. They don't need to be genuine worlds. They can be arbitrary (usually not further specified) objects whose job is to represent genuine worlds. Second,

a model must specify an “interpretation function” that tells us for each sentence letter at which of the worlds it is true.

**Definition 2.1**

A **basic model** of  $\mathcal{L}_M$  is a pair  $\langle W, V \rangle$  of

- a non-empty set  $W$ , and
- a function  $V$  that assigns to each sentence letter of  $\mathcal{L}_M$  a subset of  $W$ .

In the next chapter, we will replace this definition by a slightly more complicated definition. That’s why I’ve called models of the present kind ‘basic’.

You should be familiar with elementary concepts of set theory. A *set* is a collection of objects, called the *members* or *elements* of the set. Sets can be defined by listing their members enclosed in curly braces: ‘ $\{a, b, c\}$ ’. The *empty set*, with no members, is denoted by ‘ $\emptyset$ ’. A *subset* of a set  $X$  is a set all whose members are members of  $X$ . A *function* is a mapping – a kind of abstract machine that takes objects of a certain kind as input and outputs objects of a possibly different kind.

The interpretation function  $V$  in a model maps each sentence letter to the set of worlds at which the sentence is true. For example, if  $W$  contains three worlds  $w_1, w_2$ , and  $w_3$ , and  $V(p) = \{w_1, w_3\}$  – meaning that  $V$  maps  $p$  to the set  $\{w_1, w_3\}$  –, then  $p$  is true at  $w_1$  and  $w_3$  but not at  $w_2$ .

Notice that an interpretation function only specifies at which worlds the *sentence letters* are true.  $V$  is defined for  $p, q$ , and  $r$ , but not for  $p \rightarrow q$  or  $\Box p$  or  $\Diamond \Box q$ . This is the key idea behind the possible-worlds analysis. Once we know at which worlds each sentence letter is true, we have all we need to determine the truth-value of every sentence at every world.

To formally define how the truth-value of complex sentences is determined, I will use (meta-linguistic) statements of the form

$$M, w \models A$$

as shorthand for

$A$  is true at world  $w$  in model  $M$ .

I use ‘ $M, w \not\models A$ ’ for the negation of ‘ $M, w \models A$ ’.

Yes, it’s the same turnstile that we use for entailment and validity. This should cause no confusion because it is usually clear if the things to the left of the turnstile are  $\mathcal{L}_M$ -

sentences or meta-linguistic expressions for a model and a world. (In its present use, the turnstile is often pronounced ‘makes true’ or ‘satisfies’.)

The relation  $\models$  between a model, a world and an  $\mathcal{L}_M$ -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  $W$ ,  $P$  is any sentence letter, and  $A, B$  are any  $\mathcal{L}_M$ -sentences, then

- (a)  $M, w \models P$       iff  $w$  is in  $V(P)$ .
- (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  $M, w \not\models A$  or  $M, w \models B$ .
- (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’s go through the clauses in this definition.

Clause (a) says that a sentence letter is true at a world in a model iff the world is an element of the set of worlds which the model’s interpretation function assigns to the sentence letter. This is just what I explained above.

Clause (b) says that the negation  $\neg A$  of an  $\mathcal{L}_M$ -sentence  $A$  is true at a world in a model iff  $A$  is not true at that world in that model. In other words, 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.

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  $\mathcal{L}_M$ . (The non-logical expressions in  $\mathcal{L}_M$  don’t have a fixed meaning.)

Since every  $\mathcal{L}_M$ -sentence is built up from sentence letters with the operators covered in definition 2.2, the definition settles the truth-value of every sentence at every world in

every model.

Consider, for example, the following model  $M$ :

$$\begin{aligned} W &= \{w_1, w_2\} \\ V(p) &= \{w_1, w_2\} \\ V(q) &= \{w_1\} \\ V(P) &= \emptyset \text{ for all other sentence letters } P \end{aligned}$$

This model contains only two worlds,  $w_1$  and  $w_2$ . The interpretation function  $V$  says that  $p$  is true at both worlds,  $q$  is true at  $w_1$ , and all other sentence letters are true nowhere. 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  $w_1$  and  $w_2$  (by clause (a) of definition 2.2).
2.  $q$  is true at  $w_1$  and not true at  $w_2$  (by clause (a) of definition 2.2).
3.  $\Box p$  is true at  $w_1$  and  $w_2$  (by 1 and clause (g) of 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  $w_1$  and  $w_2$  (by 3 and clause (h) of definition 2.2).
6.  $(\Box q \rightarrow \Diamond\Box p)$  is true at  $w_1$  and  $w_2$  (by 4, 5, and clause (e) of definition 2.2).
7.  $\Diamond(\Box q \rightarrow \Diamond\Box p)$  is true at  $w_1$  and  $w_2$  (by 6 and clause (h) of definition 2.2).
8.  $\Box\Diamond(\Box q \rightarrow \Diamond\Box p)$  is true at  $w_1$  and  $w_2$  (by 7 and clause (g) of definition 2.2).

### Exercise 2.3

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

## 2.3 Basic entailment and validity

Using the concept of a model, we can sharpen the hand-wavy definitions of entailment and validity from section 1.3.

Imagine a list of all conceivable scenarios and all possible interpretations of the sentence letters. By definition 1.3, a sentence is valid iff it is true in all of these scenarios under each of these interpretations. Every combination of a scenario  $S$  and an interpretation  $I$  is represented by a model. The model contains enough information to figure out whether any given sentence is true or false in  $S$  under  $I$ . Assuming that, conversely, every model represents some combination of a scenario and an interpretation, it follows

that a sentence is valid iff it is true in every model. In the same way, some sentences  $\Gamma$  entail a sentence  $A$  iff  $A$  is true in every model in which all members of  $\Gamma$  are true.

That's the idea. There is, however, a small problem. Take a model with two worlds,  $W = \{w_1, w_2\}$ , and assume that  $V(p) = \{w_1\}$ . Is  $p$  true in this model? We can't say. Definition 2.2 only specifies under what conditions a sentence is true *at a world in a model*. We have not defined what it means for a sentence to be true in a model. So we can't say that a sentence is valid iff it is true in all models.

There are two ways to fix this. The conceptually cleaner response is to change the definition of a model. Intuitively, the worlds in a scenario are not all on a par. Think of a scenario in which it is raining although it might have been snowing. This scenario has worlds at which it is raining and others at which it is snowing. One of these worlds – a rain world – is special: it represents the actual world in the scenario. 'It is raining' is true in the scenario because it is raining in the actual world of the scenario. Following this line of thought, we could define a model to consist of *three* elements: a set of worlds  $W$ , an interpretation function  $V$ , and a "designated element of  $W$ " that indicates which world in  $W$  represents the actual world of the scenario. We could then say that a sentence is *true in a model* iff it is true at the actual world of the model. Models of this type – with a designated element of  $W$  – are called *pointed models*.

We will adopt the more popular second response. Here we change the definition of entailment and validity. Instead of saying that a sentence is valid iff it is true in every model, we say that a sentence is valid iff it is true *at every world in every model*. Similarly, we say that some sentences  $\Gamma$  entail a sentence  $A$  iff  $A$  is true at every world in every model at which all members of  $\Gamma$  are true.

The two responses amount to the same thing. Since every world in every basic (un-pointed) model could be chosen as the designated world, a sentence is true at all worlds in all basic models iff it is true in all pointed models. The response we adopt has the minor advantage of keeping models slightly simpler, and logicians want their models to be as simple as possible.

### Definition 2.3

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

### Definition 2.4

Some sentences  $\Gamma$  (**logically**) **entail** a sentence  $A$  (for short:  $\Gamma \models A$ ) iff there is no world in any basic model at which all sentences in  $\Gamma$  are true while  $A$  is false.

### Exercise 2.4

Call a sentence true *throughout* a model iff it is true at every world in the model. What do you think of the following definition? ‘ $\Gamma \models A$  iff there is no model throughout which all sentences in  $\Gamma$  are true and throughout which  $A$  is false.’ Is this equivalent to definition 2.4? (Hint: consider the hypothesis that  $p \models \Box p$ .)

Above I mentioned an assumption implicit in our new definitions: that every model represents a pair of a conceivable scenario and interpretation. This isn’t obvious. For example, if our topic is metaphysical possibility and necessity, it may be hard to conceive of a scenario with exactly two possible worlds. Is it really conceivable that there are only two ways a world might have been, compatible with the nature of things? We could stipulate that a model, at least for this application, must contain at least (say) a million worlds, or infinitely many. It turns out, however, that this would make no difference to the logic. The very same sentences are valid whether we impose the restriction or not. So we’ll allow for models with very few worlds. Such models are often useful as toy models to illustrate facts about entailment and validity.

## 2.4 Explorations in S5

By definition 2.3, a sentence is valid iff it is true at all worlds in all (basic) models. Definition 2.1 explains what a (basic) model is; definition 2.2 specifies the truth-value of any sentence at any world in any model. Together, these definitions settle which sentences are valid.

Take, for instance,  $\Box p \rightarrow p$ . This is valid on our definitions. To see why, let  $w$  be an arbitrary world in an arbitrary model  $M$ . Either  $p$  is true at  $w$  or not. If  $p$  is true at  $w$ , then by clause (e) of definition 2.2,  $\Box p \rightarrow p$  is also true at  $w$ . If  $p$  is not true at  $w$ , then by clause (g) of definition 2.2,  $\Box p$  is not true at  $w$  in  $M$ , and then  $\Box p \rightarrow p$  is true at  $w$  by clause (e). Either way,  $\Box p \rightarrow p$  is true at  $w$ . Since  $w$  and  $M$  were chosen arbitrarily, this shows that  $\Box p \rightarrow p$  is true at every world in every model.

(In the previous chapter, I mentioned that for some applications of modal logic, we

don't want  $\Box p \rightarrow p$  to be valid. In the next chapter, we will see how this can be achieved, by adding a slight tweak to the definitions of the present chapter.)

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

### Exercise 2.5

Show that  $\Box p \rightarrow \Diamond p$  is valid.

There is a shorter way to show that  $\Box p \rightarrow \Box \Box p$  is valid. Definition 2.2 entails that if a sentence starts with a modal operator, then its truth-value never varies from world to world. 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. It follows that if a sentence starts with a modal operator, then its truth-value doesn't change if you stack further modal operators in front. If  $\Diamond p$  is true at a world in a model, then so are  $\Box \Diamond p$  and  $\Diamond \Diamond p$ .

This means that any sentence that begins with a sequence of modal operators is equivalent to the same sentence with all but the last operator removed.  $\Diamond \Box \Box \Diamond p$  is equivalent to  $\Diamond p$ .  $\Box \Box p$  is equivalent to  $\Box p$ . Since replacing logically equivalent sentences inside a larger sentence never affects the larger sentence's truth-value at any world,  $\Box \Box p \rightarrow \Box p$  is equivalent to  $\Box p \rightarrow \Box p$ . And this is obviously valid.

Do not 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  $\mathcal{L}_M$ -sentence is necessary generally varies from model to model. In a model whose interpretation function makes  $p$  true at all worlds,  $p$  is necessary insofar as  $\Box p$  is true at all worlds. In a model whose interpretation function makes  $p$  false at some world,  $\Box p$  is false 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 2.6

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

Here is an example of an invalid sentence:

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

How could we show that this is invalid? By definition 2.3, 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 sentence is false. Such a model is called a **countermodel** for the sentence. The following model is a countermodel for the sentence above, as you should verify with the help of definition 2.2.

$$W = \{w, v\}$$

$$V(p) = \{w\}$$

$$V(q) = \{v\}$$

I haven't explained at which worlds sentence letters other than  $p$  and  $q$  are true, because it doesn't matter.

#### Exercise 2.7

Show that  $p \rightarrow \Box p$  is invalid (and thus  $p \not\models \Box p$ ), by giving a countermodel. Explain why this doesn't contradict the previous exercise.

#### Exercise 2.8

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

Earlier in this section, I showed that  $\Box p \rightarrow p$  and  $\Box p \rightarrow \Box \Box p$  are valid. The arguments I gave easily generalise to other sentences in place of  $p$ . So all instances of the following schemas are valid:  $\Box A \rightarrow A$  and  $\Box A \rightarrow \Box \Box A$ .

You may remember these schemas as the schemas (T) and (4) from section 1.5. You may also remember that I defined the system S5 by stipulating that it contains all in-

stances of the following schemas:

- (Dual)  $\neg\Diamond A \leftrightarrow \Box\neg A$
- (T)  $\Box A \rightarrow A$
- (K)  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
- (4)  $\Box A \rightarrow \Box\Box A$
- (5)  $\Diamond A \rightarrow \Box\Diamond A$

You can check that all instances of these schemas are valid by the definitions of the present chapter.

I also specified two rules for S5. The first says that any truth-functional consequence of any sentences in S5 is itself in S5. The second says that whenever a sentence  $A$  is in S5, then so is  $\Box A$ . As we will show chapter 4, these rules preserve validity (as defined in the previous section). Indeed, you will learn how to show that the sentences that are valid by our present definitions are precisely the sentences in S5.

In the meantime, let's prove a simpler fact to which I have appealed above (as well as on page 17 in the previous chapter): that replacing logically equivalent sentences inside a larger sentence never affects the larger sentence's truth-value at any world.

To show this, I am going to use a technique called **induction on complexity**. It works like this. Suppose we want to show that every sentence of a language has a certain property. To do so, we first show that all simple, atomic sentences of the language have the property. The atomic sentences of  $\mathcal{L}_M$  are the sentence letters. In a second step, we then show that the logical operators preserve the property, meaning that if an operator (like ' $\wedge$ ' or ' $\Box$ ') is applied to one or more sentences, and these sentences have the property, then the resulting sentence (that we get by applying the operator) still has the property. In this second step, we therefore *assume* that the sentences to which the operator is applied have the property. This is called the *induction hypothesis*. Based on this assumption, we show that the more complex resulting sentence still has the property.

**Observation 2.1:** If  $A$  is an  $\mathcal{L}_M$ -sentence and  $A'$  results from  $A$  by replacing a subsentence of  $A$  with a logically equivalent sentence, then  $A$  and  $A'$  are logically equivalent.

*Proof.* Remember that two sentences are logically equivalent if each entails the other. By definition 2.4, this means that the two sentences are true at the same worlds in every model.

Now let  $A$  be an arbitrary  $\mathcal{L}_M$ -sentence and assume that  $A'$  results from  $A$  by replacing a subsentence of  $A$  with a logically equivalent sentence. To show that  $A$  and  $A'$  are equivalent, we first consider the case where  $A$  is a sentence letter. In this case,  $A$  has no sentences as proper parts and the target hypothesis is vacuously true: there is no way of turning  $p$  into a non-equivalent sentence by replacing a subsentence within  $p$ .

Next we consider the case where  $A$  is a complex sentence that results by applying some logical operator to one or more simpler sentences. We assume (as our induction hypothesis) that the target hypothesis holds for the simpler sentences.

Assume that  $A$  is the negation of another sentence  $B$ . So  $A$  is  $\neg B$  and  $A'$  is  $\neg B'$  for some sentence  $B'$  that is either equivalent to  $B$  (if  $B$  is the subsentence of  $A$  that has been replaced to yield  $A'$ ) or that results from  $B$  by replacing a subsentence within  $B$  by an equivalent sentence (if the subsentence of  $A$  that has been replaced to yield  $A'$  isn't  $B$ ). In the latter case, our assumption that the observation holds for sentences simpler than  $A$  implies that  $B$  and  $B'$  are equivalent. Either way, then,  $B$  and  $B'$  are logically equivalent: they are true at the same worlds in every model. By clause (b) of definition 2.2, it follows that  $A$  and  $A'$  are also true at the same worlds in every model.

Essentially the same reasoning applies in the case where  $A$  is a conjunction  $B \wedge C$ , a disjunction  $B \vee C$ , a conditional  $B \rightarrow C$ , a biconditional  $B \leftrightarrow C$ , a box sentence  $\Box B$ , and a diamond sentence  $\Diamond B$ . I won't bore you by going through all of them. Here is the case for  $\Box B$ .

Assume that  $A$  has the form  $\Box B$ . So  $A$  is  $\Box B$  and  $A'$  is  $\Box B'$  for some sentence  $B'$  that is equivalent to  $B$  (by the same reasoning as before). By clause (g) of definition 2.2 it follows that  $A$  and  $A'$  are also equivalent.  $\square$

## 2.5 Trees

I will now introduce a streamlined method for working through definition 2.2 to check whether a sentence is valid: the method of **analytic tableau** or **tree proofs**. (You may be familiar with this method for non-modal logic. If so, good. If not, no problem.) It is best introduced by example.

Let's check if  $\Diamond p \rightarrow \Box p$  is valid. We do this by trying to construct a countermodel. A countermodel for  $\Diamond p \rightarrow \Box p$  is a model in which there is some world  $w$  at which  $\Diamond p \rightarrow \Box p$  is false. We start our construction by assuming that the *negation* of  $\Diamond p \rightarrow \Box p$  is *true* at  $w$ . We write this down as follows.

$$1. \quad \neg(\Diamond p \rightarrow \Box p) \quad (w) \text{ (Ass.)}$$

‘1.’ and ‘(Ass.)’ are for book-keeping; ‘Ass.’ is short for ‘Assumption’, since we’re assuming that  $\neg(\Diamond p \rightarrow \Box p)$  is true at  $w$ . Now we unfold this assumption in accordance with definition 2.2. The definition tells us that a conditional  $A \rightarrow B$  is false at a world  $w$  iff the antecedent  $A$  is true at  $w$  and the consequent  $B$  is false at  $w$ . So the assumption on line 1 implies that  $\Diamond p$  is true at  $w$  and that  $\Box p$  is false at  $w$ . We expand our “tree” (or “tableau”) by adding these consequences.

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

I have ticked off line 1 (with ‘✓’) to mark that we won’t need to look at it again. All the information in line 1 is contained in lines 2 and 3. The parenthetical ‘(1)’ at lines 2 and 3 reminds us that these lines are derived from line 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) (Ass.) ✓
2.  $\Diamond p$  (w) (1) ✓
3.  $\neg\Box p$  (w) (1)
4.  $p$  (v) (2)

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

Now the only unprocessed lines are hypotheses about sentence letters and negations of sentence letters. Sentence letters don’t have (non-trivial) subsentences, so we can’t use definition 2.2 to further break down 4 or 5. The tree is complete. We have found a countermodel for  $\Diamond p \rightarrow \Box p$ .

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

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

1.  $\neg(\Box(p \rightarrow q) \rightarrow (p \rightarrow \Box q))$  (w) (Ass.) ✓
2.  $\Box(p \rightarrow q)$  (w) (1)
3.  $\neg(p \rightarrow \Box q)$  (w) (1)

Line 1 assumes that the negation of the conditional is true at some world  $w$ . Lines 2 and 3 break down this assumption, using the fact that  $\neg(A \rightarrow B)$  is true (at a world) iff  $A$  is true and  $B$  false. We could deal with line 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)

Line 5 tells us that  $\Box q$  is false at  $w$ . We can infer that there is a world – call it  $v$  – at which  $q$  is false.

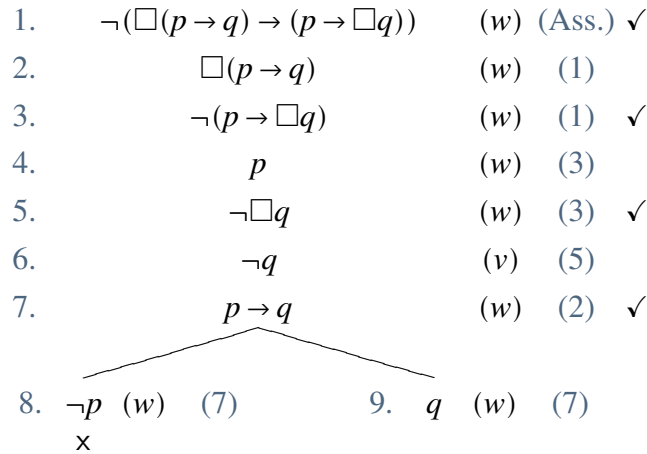
6.  $\neg q$  (v) (5)

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

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

If you are used to proofs in the natural deduction style, you may now be tempted to apply *modus ponens* and infer that  $q$  is true at  $w$ , from lines 4 and 7. In the tree method, however, we try not to draw inferences from multiple premises. We simply look at any lines that can still be processed and check what definition 2.2 tells us about the immediate subsentences of the sentence on that line. So we process line 7 without looking at line 4.

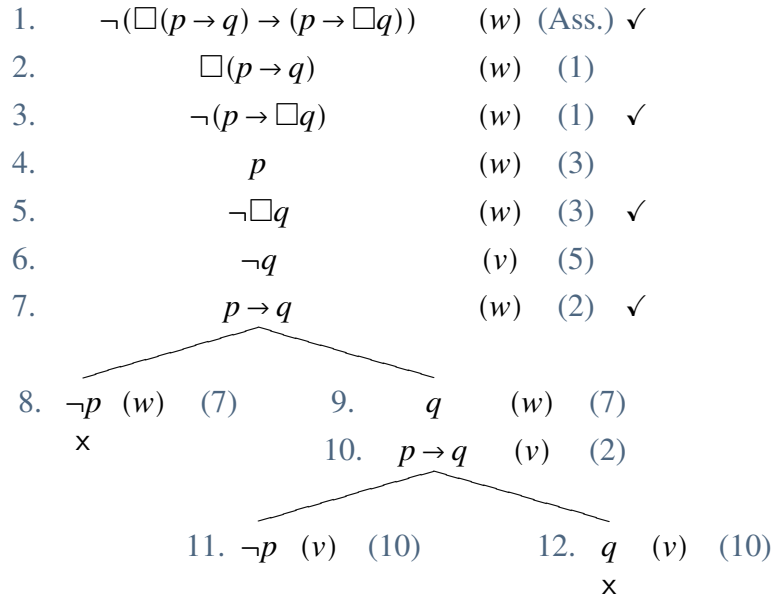
What can we infer from the truth of  $p \rightarrow q$  at  $w$  about the subsentences  $p$  and  $q$ ? By definition 2.2,  $p \rightarrow q$  is true at  $w$  if  $p$  is false at  $w$  or  $q$  is true at  $w$ . We have to keep track of both possibilities. Our (upside down) tree will branch. Here is the full tree at its present stage.



So far, I have called the numbered items on a tree ‘lines’. The proper term is **nodes**. Since nodes 8 and 9 are visually on the same line, it would be confusing to call them lines. While we’re at it, a **branch** of a tree is series of nodes that extends from the top (or “root”) node all the way down to a node below which there is no other node. The present tree has two branches, both of which contain 8 nodes.

What does this tree tell us? Remember that our aim is to construct a model in which the sentence at node 1 is true at world  $w$ . At this stage, the tree tells us that this model contains two worlds  $w$  and  $v$ ; nodes 4 and 6 tell us something about the model’s interpretation function:  $p$  is true at  $w$ ,  $q$  is false at  $v$ . After node 7, the tree branches. This means that there are two ways of extending the model we have construed so far. On the left branch, we explore an extension of the model in which  $p$  is false at  $w$ . On the right branch, we explore an extension in which  $q$  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 left 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 nodes to unfold. We have found a countermodel.

The countermodel is given by all the nodes *on the middle branch*, the one that remained open. (The other branches were dead-ends and can be ignored.) We have two worlds,  $W = \{w, v\}$ . The interpretation function  $V$  makes  $p$  true at  $w$  (node 4) and false at  $v$  (node 11);  $q$  is also true at  $w$  (node 9) and false at  $v$  (node 6). Again, you may verify that the sentence on node 1 is true at world  $w$  in this model.

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) (Ass.)

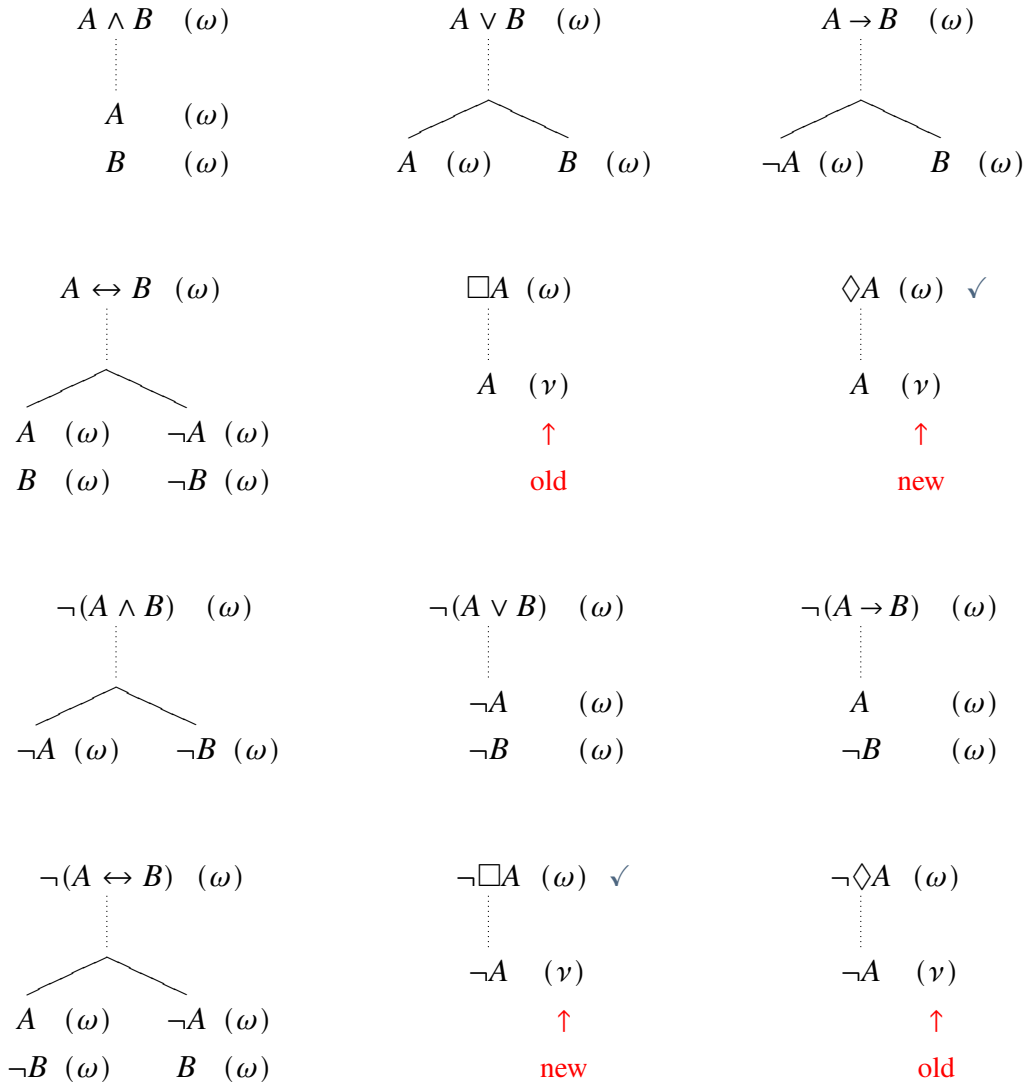
You then expand this node, and you continue expanding new nodes that appear on the tree, until no more nodes can be expanded.

To expand a node with a *non-negated sentence*, you consider what the truth of that sentence at the node's world implies for the truth-value of the sentence's immediate parts. The result may be added to the end of any open branch containing the node.

(The immediate parts of a sentence of the form  $A \wedge B$ ,  $A \vee B$ ,  $A \rightarrow B$ , or  $A \leftrightarrow B$  are the corresponding sentences  $A$  and  $B$ ; the only immediate part of  $\Box A$ ,  $\Diamond A$ , and  $\neg A$  is  $A$ .)

To expand a node with a negation  $\neg A$ , you consider what the falsity of the relevant sentence  $A$  at the node's world implies for the immediate parts of  $A$ . The result may again be added to the end of any open branch containing the node.

The following diagrams summarize how the different kinds of nodes are expanded. I use ' $\omega$ ' and ' $\nu$ ' as placeholders for arbitrary world variables.



$$\begin{array}{c} \neg\neg A \quad (\omega) \\ \vdots \\ A \quad (\omega) \end{array}$$

If a branch of a tree contains a sentence  $A$  as well as its negation  $\neg A$ , for the same world  $\omega$ , then the branch is *closed* with an  $\times$  at the bottom.

The rule for  $\Box A$  says that from the assumption that  $\Box A$  is true at a world  $\omega$  you may infer that  $A$  is true at any “old” world  $v$ , by which I mean any world *that already occurs on the branch to which you want to add a node*. You’re not allowed to introduce a new world variable ( $v$ ,  $u$ , etc.) when expanding  $\Box A$  nodes. The same is true for  $\neg\Diamond A$  nodes (which by duality means the same as  $\Box\neg A$ ). When you expand a  $\Diamond A$  node (or a  $\neg\Box A$  node), by contrast, you must introduce 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 have expanded a node that is not of type  $\Box A$  or  $\neg\Diamond A$ , and you have added the new nodes to every open branch containing the node, then you can tick off the node. You don’t need to look at it again. Nodes of type  $\Box A$  and  $\neg\Diamond A$  are never ticked off.

I have added a checkmark next to the rules for  $\Diamond A$  and  $\neg\Box A$  as a reminder that these rules can only applied once on each open branch. If you’ve already introduced a new world by expanding a  $\Diamond A$  or  $\neg\Box A$  node, you’re not allowed to introduce further new worlds on the same branch by expanding the same node again.

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

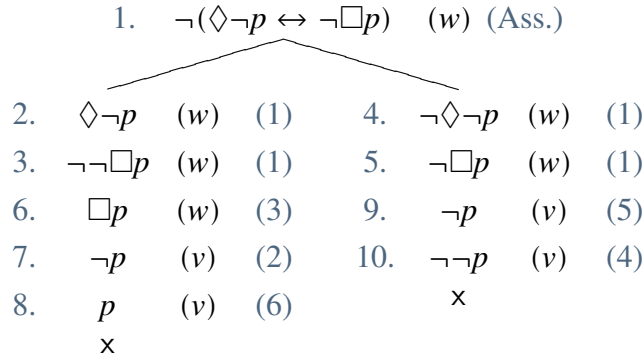
### Exercise 2.9

Use the tree method to find countermodels for the following sentences. (Spell out the countermodel, in addition to drawing the tree.)

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

What if all branches on 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 a sentence is valid.

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 all instances of a schema are valid, we can also directly draw **schematic trees** in which we use schematic variables 'A', 'B', 'C' instead of sentence letters.

### Exercise 2.10

Use the tree method to show that all instances of the following schemas are valid.

- (K)  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
- (T)  $\Box A \rightarrow A$
- (4)  $\Box A \rightarrow \Box\Box A$
- (5)  $\Diamond A \rightarrow \Box\Diamond A$

### Exercise 2.11

For each of the following sentences, either show that it is valid or give a counter-model to show that it is invalid, using the tree method.

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

(f)  $\Box\Diamond p \rightarrow \Diamond\Box p$

When constructing a tree, you often have a choice of which node to expand next. In that case, a good idea is to start with any  $\Diamond A$  or  $\neg\Box A$  nodes. If there are none, choose a node of type  $A \wedge B$ ,  $\neg(A \vee B)$  or  $\neg(A \rightarrow B)$ . Choose a node of another type only if none of the above are available. This heuristic often helps to keep trees small, but it is not part of the official tree rules.

**Exercise 2.12**

Can we use the tree method to show that some premises  $A_1, \dots, A_n$  entail a conclusion  $B$ ? Can we use it to show that two sentences  $A$  and  $B$  are equivalent?

## 3 Accessibility

### 3.1 Variable modality

I have given two kinds of interpretation for the box and the diamond. First I said that we use  $\Box A$  to express that  $A$  is certain, or historically necessary, or obligatory, etc. Then I said that  $\Box A$  means that  $A$  is true at all possible worlds. These were not meant to be competing interpretations. Rather, I have assumed that the ordinary concepts of certainty, historical necessity, obligation, and so on, can be analysed – at least to some approximation – as universal quantifiers over possible worlds.

Take historical necessity. Informally,  $A$  is historically necessary if there is nothing we can do that might render  $A$  false. Let's say that a world is *open* if there is something we can do that might bring about that world (in the sense that if we were to perform the relevant action, then we might live in that world). Now, if you think about it,  $A$  is historically necessary iff  $A$  is true at all open worlds. So we can use  $\Box A$  to mean both that  $A$  is historically necessary and that  $A$  is true at all possible worlds, where 'possible' means 'open'.

Other flavours of modality are associated with other domains of worlds. Suppose the box formalizes 'it is obligatory that'. To a first approximation,  $A$  is obligatory (relative to some norms) iff it is true at all worlds at which the norms are respected. (Think about it!) These worlds are also called *ideal*. So we can use  $\Box A$  to mean both that  $A$  is obligatory and that  $A$  is true at all possible worlds, where 'possible' now means 'ideal'.

#### Exercise 3.1

Try to analyse the following concepts in terms of universal quantification over possible worlds of a suitable kind.

- (a) It is physically necessary that  $A$ .
- (b) We know that  $A$ .
- (c) It is true that  $A$ .

Unfortunately, there are good reasons to think that this approach won't always work.

Remember that the possible-worlds semantics from the previous chapter determines a particular logic: S5. And this logic is not appropriate for every application of modal logic. In deontic logic, for example, we don't want the schema

$$(T) \quad \Box A \rightarrow A$$

to be valid. We can easily conceive of scenarios in which  $\Box p$  is true (on some interpretation of  $p$ ) even though  $p$  is false.

The semantics from the previous chapter renders the (T)-schema valid. Whenever a sentence  $\Box A$  is true at a world  $w$  in a model then  $A$  is true at  $w$  as well, because the box quantifies over all worlds, including  $w$ . To make room for deontic logic, we need a semantics in which not all worlds in  $W$  are among the “possible” worlds over which the modal operators quantify. Not all worlds are ideal.

We might also want to allow that the worlds over which the modal operators quantify depend on the world at which the relevant sentence is evaluated. Perhaps you are obligated to do the dishes in worlds where you have promised to do the dishes, but not in worlds where you haven't made the promise. Worlds in which you don't do the dishes are then ideal relative to the second kind of world, but not relative to the first.

This type of variability is also needed for other flavours of modality. Suppose the box quantifies over all worlds that are compatible with our knowledge. Which worlds are compatible with our knowledge depends on what we know. But we don't always know what we know. Sometimes we believe that we know something, but don't actually know it because it is false. We don't know it, without knowing that we don't know it. Among the worlds compatible with our knowledge are then worlds in which we know more than we actually do. What's compatible with our knowledge in *these* worlds is different from what's compatible with our knowledge in the actual world.

Let's assume, then, that for any world in any scenario there is a set of worlds that are possible *relative to*  $w$ . We assume that  $\Box p$  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*  $v$ .

Accessibility means different things in different applications. In epistemic logic, a world  $v$  is accessible from  $w$  iff  $v$  is compatible with what is known at  $w$ . In the logic of historical necessity,  $v$  is accessible from  $w$  iff  $v$  can be brought about at  $w$ . And so on.

Since facts about accessibility matter to the truth-value of modal sentences, they must be represented by our models. From now on, a model for  $\mathcal{L}_M$  will therefore specify which worlds in  $W$  are accessible from which others (and from themselves). This marks the difference between a “basic model” and a “Kripke model” – named after Saul Kripke,

who popularised models of this kind.

**Definition 3.1**

A **Kripke model** of  $\mathcal{L}_M$  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  $\mathcal{L}_M$  a subset of  $W$ .

$R$  is the accessibility relation. It is called a relation “on  $W$ ” because it holds between members of  $W$ . We write ‘ $wRv$ ’ to express that  $R$  holds between  $w$  and  $v$ .

We also need to update definition 2.2, which settles under what conditions an  $\mathcal{L}_M$ -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.

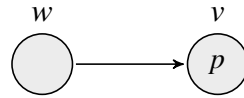
**Definition 3.2: Kripke Semantics**

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

- (a)  $M, w \models P$  iff  $w$  is in  $V(P)$ .
- (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  $M, w \not\models A$  or  $M, w \models B$ .
- (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.

To see definition 3.2 in action, consider a simple model with two worlds,  $w$  and  $v$ . World  $v$  is accessible from world  $w$ , but  $w$  is not accessible from  $v$ . Neither world can access itself. The interpretation function assigns  $\{v\}$  to  $p$  and the empty set  $\emptyset$  to all other sentence letters. The model can be pictured as follows, with an arrow representing accessibility:



Using definition 3.2, we can figure out which  $\mathcal{L}_M$ -sentences are true at which worlds in the model. For example:

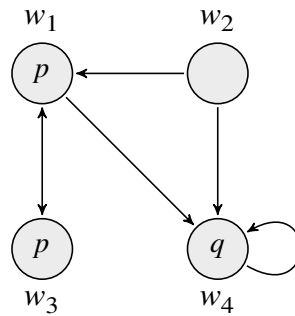
- By clause (a) of definition 3.2,  $p$  is true at  $v$  and false at  $w$ .
- By clause (h),  $\Diamond p$  is true at  $w$  because  $p$  is true at  $v$  and  $v$  is accessible from  $w$ .  $\Diamond p$  is false at  $v$  because there is no world accessible from  $v$  at which  $p$  is true.
-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

Note that  $\Diamond p$  and  $\Box \Diamond p$  have different truth-values at  $w$  (and at  $v$ ). In the new semantics, we can no longer ignore all but the last in a string of modal operators. Note also that  $\Box p$  is true at  $w$  even though  $p$  is false;  $\Box p \rightarrow p$  is no longer valid.

### Exercise 3.2

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

The next three exercises refer to the following model:



### Exercise 3.3

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 3.4

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

### Exercise 3.5

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

## 3.2 The systems K and S5

As in the previous chapter, we call a sentence *valid* if it is true at all worlds in all models. But we now use a different conception of models, and a different definition of truth at a world in a model. To avoid confusion, it is best to use different expressions for different kinds of validity. Let's call the new kind of validity *K-validity*. ('K' for Kripke.) 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 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. Entailment in the old sense (definition 2.4) will henceforth be called *S5-entailment*. Our new definition of models and truth lead to the concept of *K-entailment*.

**Definition 3.4**

Some sentences  $\Gamma$  **K-entail** a sentence  $A$  (for short:  $\Gamma \models_K A$ ) iff there is no world in any Kripke model at which all sentences in  $\Gamma$  are true while  $A$  is false.

The set of K-valid sentences is a system of modal logic. This system did not figure in C.I. Lewis's list of systems. It is known as **system K**.

K is **weaker** than S5, by which we mean that not all S5-valid sentences are K-valid.  $\Box p \rightarrow p$ , for example, is S5-valid but not K-valid. Conversely, however, every K-valid sentence is S5-valid. Let's prove this.

**Observation 3.1:** Every K-valid sentence is S5-valid.

*Proof:* In essence, observation 3.1 holds because the basic models from the previous chapter can be simulated by Kripke models in which all worlds have access to all worlds. If a sentence  $A$  is K-valid, meaning that  $A$  is true throughout every Kripke model, then  $A$  is true throughout every Kripke model of this kind, and so  $A$  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  $W$ . That is, 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, at the same world  $w$  in  $M^*$ . And if  $A$  is false at some world in some Kripke model, then  $A$  is not K-valid. By contraposition, it follows that if  $A$  is K-valid, then  $A$  is S5-valid.  $\square$

You may remember from section 1.5 that S5 can be axiomatized by five axiom schemas and two rules:

- (Dual)  $\neg\Diamond A \leftrightarrow \Box\neg A$
- (K)  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
- (T)  $\Box A \rightarrow A$
- (4)  $\Box A \rightarrow \Box\Box A$
- (5)  $\Diamond A \rightarrow \Box\Diamond A$
- (Nec) If  $A$  is in the system, then so is  $\Box A$ .
- (CPL) If  $\Gamma \models_0 A$  and all members of  $\Gamma$  are in the system, then so is  $A$ .

All instances of (Dual), (K), (T), (4), and (5) are S5-valid, and all and only the S5-valid sentences can be derived from instances of these axioms by (Nec) and (CPL).

The system K can be axiomatized by dropping three of the axiom schemas: (T), (4), and (5), leaving only (Dual) and (K). All and only the K-valid sentences can be derived from instances of (Dual) and (K) by (Nec) and (CPL).

(Many authors define  $\Box$  as  $\neg\Diamond\neg$  or  $\Diamond$  as  $\neg\Box\neg$ , in which case (Dual) is true by definition. The only remaining axiom schema is then (K). Don't confuse the schema (K) with the system K!)

#### Exercise 3.6

- (a) Describe a Kripke model in which some instance of (4) is false at some world.
- (b) Describe a Kripke model in which some instance of (5) is false at some world.

#### Exercise 3.7

Can you find an instance of the (T)-schema  $\Box A \rightarrow A$  that is K-valid?

#### Exercise 3.8

Show that  $\Box(p \vee \neg p)$  is K-valid, using definition 3.2.

### 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. Suppose, for example, we read the box as physical necessity

and the diamond as physical possibility, understood as compatibility with the laws of nature. On a popular conception of what it means to be a law of nature, nothing that happens is ever incompatible with the laws of nature. Equivalently, anything that is physically necessary is actually the case. We therefore want  $\Box A$  to 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? Below I will argue that we can't. If that is right, then the logic of physical necessity is neither K nor S5. We want a logic with (T) ( $\Box A \rightarrow A$ ) but without (4) ( $\Box A \rightarrow \Box\Box A$ ). S5 gives us both, K gives us neither.

Our current semantics makes it easy to define systems in between K and S5 by putting restrictions on the accessibility relation in Kripke models.

Let's say that an  $\mathcal{L}_M$ -sentence is **valid in a class of Kripke models** iff the sentence is true at every world in every model that belongs to the class. K-validity is validity in the class of all Kripke models. S5-validity is validity in the class of Kripke models in which every world has access to every world (as mentioned earlier, in the proof of observation 3.1).

If you inspect countermodels to the K-validity of  $\Box p \rightarrow p$ , you may notice that all of them involve worlds that don't have access to themselves. If we require that every world can see itself then all instances of the (T)-schema become valid.

**Observation 3.2:** All instances of (T) are valid in the class of Kripke models in which every world is accessible from itself.

*Proof:* According to clause (e) of definition 3.2, an instance of  $\Box A \rightarrow A$  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.  $\square$

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 also call the model itself reflexive. Observation 3.2 therefore states that all instances of (T) are valid in the class of reflexive Kripke models.

The set of all sentences that are valid in the class of reflexive Kripke models is known as **system T**. Accordingly, any sentence that is valid in this class of Kripke models (every member of system T) is called **T-valid**.

System T is stronger than K, but weaker than S5. The system can be axiomatized by adding the axiom schema (T) to the axioms and rules of K. We don't have (4) or (5).  $\Box p \rightarrow \Box\Box p$  is S5-valid but not T-valid.

Systems of modal logic sometimes share their name with a schema. For disambiguation, I always put schema names in parentheses. (T) is a schema, T is a system. (K) is a schema, K is a system. All instances of (T) are in T, but many sentences in T – for example, all instances of (K) – are not instances of (T).

### Exercise 3.9

Show that  $\Box p \rightarrow \Diamond p$  is T-valid.

In chapter 7, we will study a temporal application of modal logic in which the box is read as ‘it is always going to be the case that’. The “worlds” in a Kripke model here represent times.  $\Box p$  is understood to be true at a time  $t$  iff  $p$  is true at all times after  $t$ . The accessibility relation is the earlier-later relation:  $t_1 R t_2$  iff  $t_1$  is earlier than  $t_2$ . In this application, 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  $t_3$ .

A relation  $R$  is called **transitive** if whenever  $xRy$  and  $yRz$  then  $xRz$ . As before, we call a Kripke model transitive if its accessibility relation is transitive. When we do temporal logic, we will restrict the relevant models to transitive models.

The set of sentences that are valid in the class of transitive Kripke models is known as **system K4**. The name alludes to the fact that this system can be axiomatized by adding schema (4) to the axioms and rules of K.

**Observation 3.3:** All instances of (4) are 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 (i)  $\Box A$  is true at  $w$  and (ii)  $\Box\Box A$  is false at  $w$ . By clause (g) of definition 3.2, (ii) 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 (i),  $A$  is true at  $u$ . So  $A$  is both true and false at  $u$ . Contradiction.  $\square$

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**. It is stronger than K, T, and K4, but weaker than S5.

There are many other conditions we could impose on the accessibility relation, and many combinations of these conditions. Each of them defines a system of modal logic. The following table lists some well-known model classes with the conventional names for the corresponding systems, repeating (for future reference) the ones we already know. We will have a closer look at some of these systems in later chapters, when we turn to applications of modal logic.

System	Constraint on $R$
K	–
T	$R$ is <b>reflexive</b> : every world in $W$ can access itself
D	$R$ is <b>serial</b> : every world in $W$ can access some world
K4	$R$ is <b>transitive</b> : whenever $wRv$ and $vRu$ , then $wRu$
K5	$R$ is <b>euclidean</b> : whenever $wRv$ and $wRu$ , then $vRu$
KD45	$R$ is serial, transitive, and euclidean
B	$R$ is reflexive and <b>symmetric</b> : whenever $wRv$ then $vRw$
S4	$R$ is reflexive and transitive
S4.2	$R$ is reflexive, transitive, and <b>convergent</b> : whenever $wRv$ and $wRu$ , then there is some $t$ such that $vRt$ and $uRt$
S5	$R$ is reflexive, transitive, and symmetric
S5	$R$ is <b>universal</b> : every world has access to every world

S5 occurs twice in the 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 merely 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 divides the members of the set into classes within which everything stands in the relation to everything. (These classes are called **equivalence classes**.)

For example, let  $S$  be the relation that holds between two people iff they have the same birthday. This is an equivalence relation. It is reflexive: everyone has the same birthday as themselves. It is transitive: if  $aSb$  and  $bSc$  then  $aSc$ . And it is symmetric: if  $aSb$  then  $bSa$ . For any person  $a$ , consider the class  $[a]_S$  of everyone who has the same birthday as  $a$ . (A “class” is essentially the same thing as a set.) 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 3.4:** 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 sketch:* 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 relation on  $W$  is a special kind of equivalence relation on  $W$ . 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 assume  $A$  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 a 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  $W'$ .

$V'$  is the restriction of  $V$  to  $W'$ , so that for any sentence letter  $B$ ,  $V'(B) = V(B) \cap W'$ .

(If  $X$  and  $Y$  are sets, then  $X \cap Y$  – the *intersection* of  $X$  and  $Y$  – is the set of all things that are both in  $X$  and in  $Y$ .)

$M'$  has a universal accessibility relation. But from the perspective of  $w$ ,  $M$  and  $M'$  are indistinguishable. *Any sentence is true at  $w$  in  $M$  iff it is true at  $w$  in  $M'$ .* This could be shown by induction, but I hope you see intuitively why it is the case.

Granting the italicized sentence, the assumption that  $A$  is false at some world in some model whose accessibility relation is an equivalence relation entails that  $A$  is false at some world in some model whose accessibility relation is universal.  $\square$

### Exercise 3.10

Let  $R$  be the relation on the set of all people such that  $aRb$  iff  $b$  is a sibling of  $a$ . Is  $R$  reflexive? serial? transitive? euclidean? symmetric? universal?

### Exercise 3.11

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 3.12

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.”

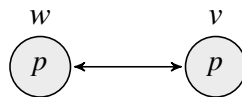
## 3.4 Frames

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. What exactly is the connection?

You might think the connection between (T) and reflexivity is this:

(?) All instances of (T) are valid in a model iff the model is reflexive.

But that’s false. We know (observation 3.2) that all (T) instances are valid in the class of reflexive models. It follows that all (T) instances are valid in every reflexive model. But the other direction fails. There are non-reflexive models in which all (T) instances are valid. The following model is an example.



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. The class might contain models like the one just described.

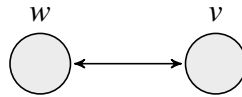
To understand the connection between modal schemas and conditions on the accessibility relation, we need to talk about *frames*. A frame is what you get if take a model and remove the interpretation function.

**Definition 3.5**

A **Kripke frame** is a pair of a non-empty set  $W$  and a relation  $R$  on  $W$ .

Roughly speaking, if we think of a model as representing a scenario and an interpretation, then a frame is the part of the model that represents the scenario.

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



Now remember that validity is 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.

To see the point, suppose I suggested that a sentence is “X-valid” iff it is true at all worlds in all Kripke model whose interpretation function assigns the empty set to the sentence letter  $p$ . So  $\Box \neg p$  is X-valid, while  $\Box \neg q$  is X-invalid. But  $\Box \neg p$  and  $\Box \neg q$  have the same logical form. If  $\Box \neg p$  is true in virtue of its logical form, then  $\Box \neg q$  should also be true in virtue of its logical form. X-validity is not a sensible concept of logical validity. The systems from the previous section were all defined sensibly, by putting constraints on the frame of a Kripke model, not on 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.

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. We could have defined the systems from

the previous section in terms of frame classes rather than model classes:  $K$  is the set of sentences valid in the class of all frames,  $T$  is the set of sentences valid in the class of reflexive 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: All  $(T)$  instances are valid in a class of frames iff every frame in the class is reflexive. More simply:

**Observation 3.5:** All instances of  $(T)$  are valid on a frame iff the frame is reflexive.

*Proof:* The right-to-left direction follows from observation 3.2, according to which all  $(T)$  instances are 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.) Let  $V(p)$  comprise all worlds in  $W$  except  $w$ . Then  $\Box p$  is true at  $w$  and  $p$  false. So  $\Box p \rightarrow p$  is false at  $w$ .  $\square$

If all instances of a schema are valid on all and only the frames whose accessibility relation satisfies a certain property, the schema is said to **correspond** to that property (and to *define* the relevant class of frames). Observation 3.5 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 from the start of section 3.3. I claimed that on this interpretation of the box and the diamond, we should not regard all instances of the (4)-schema  $\Box A \rightarrow \Box \Box A$  as valid. My claim is not based on a direct intuition that something could be physically necessary without it being physically necessary that it is physically necessary. My claim is rather based on a judgement about the non-transitivity of physical accessibility. My reasoning goes like this. I assume 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  $w$ . For example, 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  $w$ . Since (4) corresponds to transitivity, I can infer that the logic of physical necessity does not render all instances of that schema valid.

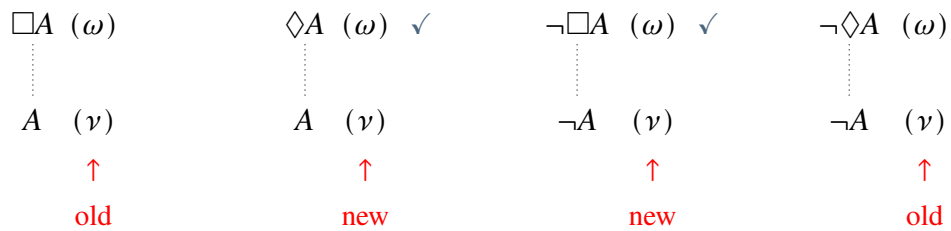
### Exercise 3.13

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 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 tree 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.

$$n. \quad \Diamond p \quad (w)$$

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:

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

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  $\mathcal{L}_M$ ; 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?

$$n. \quad \Box p \quad (w)$$

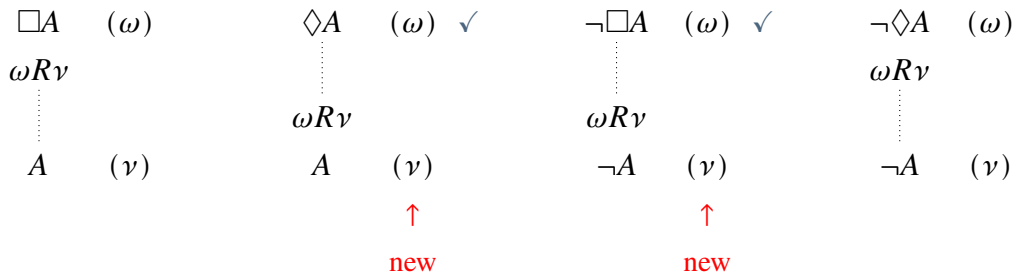
By itself, this doesn't tell us anything about the truth-value of  $p$  at any world. We can't infer that  $p$  is true at  $w$ , because  $w$  might not be accessible from itself. Indeed, if no world is accessible from  $w$ , then  $\Box p$  can be true even if  $p$  is false at every world. So we can't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

$$m. \quad wRv$$

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  $w$  at which the box sentence is true has access to some world  $v$ .

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).

As before, the checkmark next to the rules for  $\Diamond A$  and  $\neg\Box A$  indicates that these nodes can only be expanded once on each branch.

The rules for negated boxes and diamonds are what 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  $v$ .

What happens if we try to prove  $\Box p \rightarrow p$ ?

1.  $\neg(\Box p \rightarrow p)$  (w) (Ass.)
2.  $\Box p$  (w) (1)
3.  $\neg p$  (w) (1)

At this point, no more rules can be applied. We can read off a countermodel from the open branch:

$$\begin{aligned} W &= \{w\} \\ R &= \emptyset \\ V(p) &= \emptyset \end{aligned}$$

This is the smallest possible Kripke model. It consists of a single world that can't see itself. ' $R = \emptyset$ ' is a way of saying that no world can see any world. If you want to say that  $R$  holds between  $w$  and  $v$  and between  $v$  and  $u$ , you might write ' $R = \{(w, v), (v, u)\}$ ' or simply ' $wRv, vRu$ '.

### Exercise 3.14

Use the K-rules to check which of the following sentences are K-valid. If a sentence is invalid, describe a countermodel.

- (a)  $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$
- (b)  $\Diamond(p \wedge q) \rightarrow (\Diamond p \wedge \Diamond q)$
- (c)  $(\Diamond p \wedge \Diamond q) \rightarrow \Diamond(p \wedge q)$
- (d)  $\Diamond(p \vee q) \leftrightarrow (\Diamond p \vee \Diamond q)$
- (e)  $\Box(p \vee q) \leftrightarrow (\Box p \vee \Box q)$
- (f)  $\Box(p \rightarrow q) \rightarrow (\Diamond p \rightarrow \Diamond q)$ .
- (g)  $(\Box p \wedge \Diamond q) \rightarrow \Diamond(p \wedge q)$ .

For systems in between K and S5 that are characterised by certain constraints on the accessibility relation, we add new rules for manipulating accessibility nodes. For example, if we want to check whether a sentence is T-valid, we use a *reflexivity rule* in addition to the K-rules. The reflexivity rule says that if a world variable  $\omega$  occurs on a branch, then we may always add  $\omega R \omega$  to the branch.

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

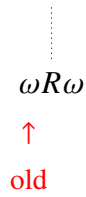
- |    |                              |            |
|----|------------------------------|------------|
| 1. | $\neg(\Box p \rightarrow p)$ | (w) (Ass.) |
| 2. | $\Box p$                     | (w) (1)    |
| 3. | $\neg p$                     | (w) (1)    |
| 4. | $wRw$                        | (Ref.)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

To test for validity in the class of transitive frames (or models), we need a *transitivity rule*, which allows us to infer  $\omega Rv$  from  $\omega Rv$  and  $vRu$ . Here is a proof of  $\Box p \rightarrow \Box \Box p$  that uses this rule.

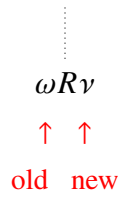
1.	$\neg(\Box p \rightarrow \Box\Box p)$	(w)	(Ass.)
2.	$\Box p$	(w)	(1)
3.	$\neg\Box\Box p$	(w)	(1)
4.	$wRv$		(3)
5.	$\neg\Box p$	(v)	(3)
6.	$vRu$		(5)
7.	$\neg p$	(u)	(5)
8.	$wRu$		(4,6,Tr.)
9.	$p$	(u)	(2,8)
	x		

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

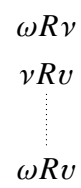
Reflexivity



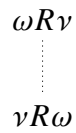
Seriality



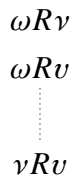
Transitivity



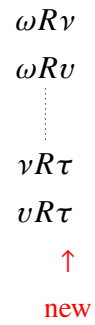
Symmetry



Euclidity



Convergence



By selectively adding some of these rules to the K-rules, we get tree rules for a variety of modal logics. (Compare the table on p. 56.)

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
KD45	K-rules, seriality rule, transitivity rule, and euclidity rule
B	K-rules, reflexivity rule, and symmetry rule
S4	K-rules, reflexivity rule, and transitivity rule
S4.2	K-rules, reflexivity rule, transitivity rule, and convergence rule

### Exercise 3.15

Use the tree method to check the following claims.

- (a)  $\models_{K4} \Diamond p \rightarrow \Diamond \Diamond p$ .
- (b)  $\models_D (\Box p \wedge \Box q) \rightarrow \Diamond(p \vee q)$ .
- (c)  $\models_B \Diamond p \rightarrow \Box \Diamond p$ .
- (d)  $\models_T (\Diamond \Box(p \rightarrow q) \wedge \Box p) \rightarrow \Diamond q$ .
- (e)  $\models_T \Diamond(p \rightarrow \Box \Diamond p)$ .
- (f)  $\models_{S4} \Diamond \Box(\Diamond p \rightarrow \Box \Diamond p)$ .

## 4 Models and Proofs

### 4.1 Soundness and completeness

You may find that this chapter is harder and more abstract than the previous chapters. Feel free to skip or skim it if you're mostly interested in philosophical applications.

We have introduced several kinds of validity: S5-validity, K-validity, T-validity, and so on. All of these are defined in terms of models. K-validity means truth at all worlds in all Kripke models. T-validity means truth at all worlds in all reflexive Kripke models. S5-validity means truth at all worlds in all universal Kripke models (equivalently, at all worlds in all “basic” models). And so on.

If you want to show that a sentence is, say, K-valid, you could directly work through the clauses of definition 3.2, showing that there is no world in any Kripke model in which the sentence is false. The tree method regiment and simplifies this process. If you construct a tree for your sentence in accordance with the K-rules and all branches close, then the sentence is K-valid. If some branch remains open, the sentence isn't K-valid.

Or so I claimed. But these claims aren't obvious. The tree rule for the diamond, for example, appears to assume that if  $\Diamond A$  is true at a world then  $A$  is true at some accessible world *that does not yet occur on the branch*. Couldn't  $\Diamond A$  be true because  $A$  is true at an accessible “old” world instead? Also, why do we expand  $\Diamond A$  nodes only once? Couldn't  $A$  be true at multiple accessible worlds?

In the next two sections, we are going to lay any such worries to rest. We are going to prove that (1) if all branches on a K-tree close then the target sentence is K-valid; conversely, (2) if some branch on a fully developed K-tree remains open, then the target sentence is not K-valid. (1) establishes the *soundness* of the tree rules for K, (2) establishes their *completeness*.

When you use the tree method, you don't have to think of what you are doing as exploring Kripke models. I could have introduced the method as a purely syntactic game. You start the game by writing down the negation of the target sentence, followed by ‘(w)’ (and possibly ‘1.’ to the left and ‘(Ass.)’ to the right, although in this chapter we will mostly ignore these book-keeping annotations.) Then you repeatedly apply the tree rules

until either all branches are closed or no rule can be applied any more. At no point in the game do you need to think about what any of the symbols you are writing might mean.

Soundness and completeness link this syntactic game with the “model-theoretic” concept of validity. Soundness says that if the game leads to a closed tree (a tree in which all branches are closed) then the target sentence is true at all worlds in all models. Completeness says that if the game doesn’t lead to a closed tree then the target sentence is not true at all worlds in all models. This is called completeness because it implies that every valid sentence can be shown to be valid with the tree method.

In general, a proof method is called **sound** if everything that is provable with the method is valid. A method is **complete** if everything that is valid is provable. Strictly speaking, we should say that a method is sound or complete *for a given concept of validity*. The tree rules for K are sound and complete for *K-validity*, but not for T-validity or S5-validity.

The tree method is not the only method for showing that a sentence is K-valid (or T-valid, or S5-valid). Instead of constructing a K-tree, you could construct an axiomatic proof, trying to derive the target sentence from some instances of (Dual) and (K) by (Nec) and (CPL). This, too, can be done as a purely syntactic exercise, without thinking about models and worlds. In section 4.4, we will show that the axiomatic calculus for K is indeed sound and complete for K-validity: all and only the K-valid sentences can be derived from (Dual) and (K) by (Nec) and (CPL). The ‘all’ part is completeness, the ‘only’ part soundness. Having shown soundness and completeness for both the tree method and the axiomatic method, we will have shown that the two methods are equivalent. Anything that can be shown with one method can also be shown with the other.

There are other styles of proof besides the axiomatic and the tree format. Two famous styles that we won’t cover are “natural deduction” methods and “sequence calculi”. Logicians are liberal about what qualifies as a proof method. The only non-negotiable condition is that there must be a mechanical way of checking whether something (usually, some configuration of symbols) is or is not a proof of a given target sentence.

#### Exercise 4.1

What do you think of the following proposals for new proof methods?

- (a) In *method A*, every  $\mathcal{L}_M$ -sentence is a proof of itself: To prove an  $\mathcal{L}_M$ -sentence with this method, you simply write down the sentence.
- (b) In *method B*, every  $\mathcal{L}_M$ -sentence that is an instance of  $\Box(A \vee \neg A)$  is a proof of itself. Nothing else is a proof in method B.

(c) In *method C*, a proof of a sentence  $A$  is a list of  $\mathcal{L}_M$ -sentences terminating with  $A$  and in which every sentence occurs in some logic textbook. Which of these qualify as genuine proof methods by the criterion I have described?

### Exercise 4.2

Which, if any, of the methods from the previous exercise are sound for K-validity? Which, if any, are complete?

## 4.2 Soundness for trees

We are now going to show that the tree method for K is sound – that every sentence that can be proved with the method is K-valid. A proof in the tree method is a tree in which all branches are closed. So this is what we have to show:

Whenever all branches on a K-tree close then the target sentence is K-valid.

By a *K-tree* I mean a tree that conforms to the K-rules from the previous chapter.

I'll first explain the proof idea, then I'll fill in the details. We will assume that there is a K-tree for some target sentence  $A$  on which all branches close. We need to show that  $A$  is K-valid. To this end, we suppose for reductio that  $A$  is *not* K-valid. By definition 3.3, a sentence is K-valid iff it is true at all worlds in all Kripke models. Our supposition that  $A$  is not K-valid therefore means that  $A$  is false at some world in some Kripke model. Let's call that world ' $w$ ' and the model ' $M$ '. Note that the closed tree begins with

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

If we take the world variable ' $w$ ' on the tree to pick out world  $w$  in  $M$ , then node 1 is a correct statement about  $M$ , insofar as  $\neg A$  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  $M$ , 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  $M$ .*

Since our closed tree is constructed from node 1 by applying the K-rules, it follows that all nodes on some branch of the tree are correct statements about  $M$ . But every branch

of a closed tree contains a pair of contradictory statements, which can't both be correct statements about  $M$ . This completes the reductio.

Let's fill in the details. We first define precisely what it means for the nodes on a tree branch to be correct statements about a model.

#### Definition 4.1

A tree node is a **correct statement about** a Kripke model  $M = \langle M, R, V \rangle$  **under** a function  $f$  that maps world variables to members of  $W$  iff either the node has the form  $\omega R v$  and  $f(\omega) R f(v)$ , or the node has the form  $A(\omega)$  and  $A$  is true at  $f(\omega)$  in  $M$ .

A tree branch **correctly describes** a model  $M$  iff there is a function  $f$  under which all nodes on the branch are correct statements about  $M$ .

We now prove the italicised statement above:

#### Soundness Lemma

If some branch  $\beta$  on a tree correctly describes a Kripke model  $M$ , and the branch is extended by applying a K-rule, then at least one of the resulting branches correctly describes  $M$ .

*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 a branch  $\beta$  that correctly describes  $M$ , so that there is a function  $f$  under which all nodes on the branch are correct statements about  $M$ . We show that once the rule has been applied, at least one of the resulting branches still correctly describes  $M$ .

- Suppose  $\beta$  contains a node of the form  $A \wedge B(\omega)$  and the branch is extended by two new nodes  $A(\omega)$  and  $B(\omega)$ . Since  $A \wedge B(\omega)$  is a correct statement about  $M$  under  $f$ , we have  $M, f(\omega) \models A \wedge B$ . By clause (c) of definition 3.2, it follows that  $M, f(\omega) \models A$  and  $M, f(\omega) \models B$ . So the extended branch still correctly describes  $M$ .
- Suppose  $\beta$  contains a node of the form  $A \vee B(\omega)$  and the branch is split into two, with  $A(\omega)$  appended to one end and  $B(\omega)$  to the other. Since the expanded node is a correct statement about  $M$  under  $f$ , we have  $M, f(\omega) \models A \vee B$ . By clause (d) of definition 3.2, it follows that either  $M, f(\omega) \models A$  or  $M, f(\omega) \models B$ . So at least one of the resulting branches also correctly describes  $M$ .

The proof for the other non-modal rules is similar. Let's look at the rules for the modal operators.

- Suppose  $\beta$  contains nodes of the form  $\Box A(\omega)$  and  $\omega Rv$ , and the branch is extended by adding  $A(v)$ . Since  $\Box A(\omega)$  and  $\omega Rv$  are correct statements about  $M$  under  $f$ , we have  $M, f(\omega) \models \Box A$  and  $f(\omega) R f(v)$ . By clause (g) of definition 3.2, it follows that  $M, f(v) \models A$ . So the extended branch correctly describes  $M$ .
- Suppose  $\beta$  contains a node of the form  $\Diamond A(\omega)$  and the branch is extended by adding nodes  $\omega Rv$  and  $A(v)$ , where  $v$  is new on the branch. Since  $\Diamond A(\omega)$  is a correct statement about  $M$  under  $f$ , we have  $M, f(\omega) \models \Diamond A$ . By clause (h) of definition 3.2, it follows that  $M, v \models A$  for some  $v$  in  $W$  such that  $f(\omega) R v$ . Let  $f'$  be the same as  $f$  except that  $f'(v) = v$ . The newly added nodes are correct statements about  $M$  under  $f'$ . Since  $v$  is new on the branch, all earlier nodes on the branch are also correct statements about  $M$  under  $f'$ . So the expanded branch correctly describes  $M$ .

The cases for  $\neg\Box$  and  $\neg\Diamond$  are similar to the previous two cases.  $\square$

With the help of this lemma, we can prove that the method of K-trees is sound.

**Theorem: Soundness of K-trees**

If a K-tree for a target sentence closes, then the target sentence is K-valid.

*Proof:* Suppose for reductio that some K-tree for some target sentence  $A$  closes even though  $A$  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  $M$  under the function that maps the world variable ' $w$ ' to  $w$ . Since the tree is created from the first node by applying the K-rules, the Soundness Lemma implies that some branch  $\beta$  on the tree correctly describes  $M$ : all nodes on the tree are correct statements about  $M$  under some function  $f$ . But the tree is closed. This means that  $\beta$  contains contradictory nodes of the form

- n.  $B(v)$
- m.  $\neg B(v)$

If both of these are correct statements about  $M$  under  $f$ , then  $M, f(v) \models B$  and also  $M, f(v) \models \neg B$ . This is impossible by definition 3.2.  $\square$

### Exercise 4.3

Spell out the cases for  $A \rightarrow B$  and  $\neg\Diamond A$  in the proof of the Soundness Lemma.

### Exercise 4.4

Draw the K-tree for target sentence  $\Box p$ . The tree has a single open branch. Does this branch correctly describe the Kripke model in which there is just one world  $w$ ,  $w$  has access to itself, and all sentence letters are false at  $w$ ?

The soundness proof for K-trees is easily adapted to other types of trees. The tree rules for system T, for example, are all the K-rules plus the Reflexivity rule, which allows adding  $\omega R \omega$  for every world  $\omega$  on the branch. Suppose we want to show that everything that is provable with the T-rules is T-valid – true at every world in every reflexive Kripke model. All the clauses in the Soundness Lemma still hold if we assume that the model  $M$  is reflexive. We only need to add a further clause for the Reflexivity rule, to confirm that if a branch correctly describes a reflexive model  $M$ , and the branch is extended by adding  $\omega R \omega$ , then the resulting branch also correctly describes  $M$ . This is evidently the case.

### Exercise 4.5

How would we need to adjust the soundness proof to show that the tree rules for K4 are sound with respect to K4-validity?

## 4.3 Completeness for trees

Let's now show that the tree rules for K are complete – that whenever a sentence is K-valid then there is a closed K-tree for that sentence. In fact, we will show something stronger:

If a sentence is K-valid, then every fully developed K-tree for the sentence is closed.

By a *fully developed* tree, I mean a tree on which every node on any open branch that can be expanded (in any way) has been expanded (in this way). A fully developed tree may be infinite.

We will prove the displayed sentence by proving its contraposition:

If a fully developed K-tree for a sentence does not close, then the sentence is not K-valid.

Assume, then, that some fully developed K-tree for some target sentence has at least one open branch. We want to show that the target sentence is false at some world in some Kripke model.

We already know how to read off a countermodel from an open branch. All we need to do is show that this method for generating countermodels really works. Let's first define the method more precisely.

#### Definition 4.2

The model **induced by** a tree branch is the Kripke model  $(W, R, V)$  where

- (a)  $W$  is the set of world variables on the branch,
- (b)  $\omega R \nu$  holds in the model iff a node  $\omega R \nu$  occurs on the branch,
- (c) for any sentence letter  $P$ ,  $V(P)$  is the set of world variables  $\omega$  for which a node  $P(\omega)$  occurs on the branch.

Next we show that all nodes on any open branch on a fully developed tree are correct statements about the Kripke model induced by the branch.

#### Completeness Lemma

Let  $\beta$  be an open branch on a fully developed K-tree, and let  $M = \langle W, R, V \rangle$  be the model induced by  $\beta$ . Then  $M, \omega \models A$  for all sentences  $A$  and world variables  $\omega$  for which  $A(\omega)$  is on  $\beta$ .

We have to show that whenever  $A(\omega)$  occurs on  $\beta$  then  $M, \omega \models A$ . The proof is by induction on the length of  $A$ . We first show that the claim holds for sentence letters and negated sentence letters. Then we show that *if* the claim holds for all sentences shorter than  $A$  (this is our induction hypothesis), *then* it also holds for  $A$  itself.

- If  $A$  is a sentence letter then the claim is true by clause (c) of definition 4.2 and clause (a) of definition 3.2.
- If  $A$  is the negation of a sentence letter  $\rho$ , then  $\rho(\omega)$  does not occur on  $\beta$ , otherwise  $\beta$  would be closed. By clause (c) of definition 4.2, it follows that  $\omega$  is not in  $V(\rho)$ , and so  $M, \omega \models A$  by clauses (a) and (b) of definition 3.2.

- If  $A$  is a doubly negated sentence  $\neg\neg B$ , then  $\beta$  contains a node  $B(\omega)$ , because the tree is fully developed. By induction hypothesis,  $M, \omega \models B$ . By clause (b) of definition 3.2, it follows that  $M, \omega \models A$ .
- If  $A$  is a conjunction  $B \wedge C$ , then  $\beta$  contains nodes  $B(\omega)$  and  $C(\omega)$ . By induction hypothesis,  $M, \omega \models B$  and  $M, \omega \models C$ . By clause (c) of definition 3.2, it follows that  $M, \omega \models A$ .
- If  $A$  is a negated conjunction  $\neg(B \wedge C)$ , then  $\beta$  contains either  $\neg B(\omega)$  or  $\neg C(\omega)$ . By induction hypothesis,  $M, \omega \models \neg B$  or  $M, \omega \models \neg C$ . Either way, clauses (b) and (c) of definition 3.2 imply that  $M, \omega \models A$ .

I skip the cases where  $A$  is a disjunction, a conditional, a biconditional, or a negated disjunction, conditional, or biconditional. The proofs are similar to one (or both) of the previous two cases.

- If  $A$  is a box sentence  $\Box B$ , then  $\beta$  contains a node  $B(v)$  for each world variable  $v$  for which  $\omega Rv$  is on  $\beta$  (because the tree is fully developed). By induction hypothesis,  $M, v \models B$ , for each such  $v$ . By definition 4.2, it follows that  $M, v \models B$  for all worlds  $v$  such that  $\omega Rv$ . By clause (g) of definition 3.2, it follows that  $M, \omega \models \Box B$ .
- If  $A$  is a diamond sentence  $\Diamond B$ , then there is a world variable  $v$  for which  $\omega Rv$  and  $B(v)$  are on  $\beta$ . By induction hypothesis,  $M, v \models B$ . And by definition 4.2,  $\omega Rv$ . By clause (h) of definition 3.2, it follows that  $M, \omega \models \Diamond B$ .

For the case where  $A$  has the form  $\neg\Box B$  or  $\neg\Diamond B$ , the proof is similar to one of the previous two cases.  $\square$

To establish completeness, we need to verify one more point: that one can always construct a fully developed tree for any invalid target sentence. Let's call a K-tree *regular* if it is constructed by (i) first applying all rules for the truth-functional connectives until no more of them can be applied (without adding only nodes to a branch that are already on the branch), then (ii) applying the rules for  $\Diamond$  and  $\neg\Box$  until no more of them can be applied, then (iii) applying the rules for  $\Box$  and  $\neg\Diamond$  until no more of them can be applied, then starting over with (i), and so on.

**Observation 4.1:** Every regular open K-tree is fully developed.

*Proof:* When constructing a regular tree, every iteration of (i), (ii), and (iii) only allows expanding finitely many nodes. So every node on every open branch that can be expanded in any way is eventually expanded in this way by some iteration of (i), (ii), and (iii).  $\square$

Now we have all the ingredients to prove completeness.

**Theorem: Completeness of K-trees**

If a sentence is K-valid, then there is a closed K-tree for that sentence.

*Proof:* Let  $A$  be any K-valid sentence, and suppose for reductio that there is no closed K-tree for  $A$ . In particular, then, every regular K-tree for  $A$  remains open. Take any such tree. By observation 4.1, the tree is fully expanded. Choose any open branch on the tree. By the Completeness Lemma,  $A$  is false at  $w$  in the model induced by that branch. So  $A$  is not true at all worlds in all Kripke models. Contradiction.  $\square$

**Exercise 4.6**

Fill in the cases for  $B \rightarrow C$  and  $\neg\Diamond B$  in the proof of the Completeness Lemma.

Like the soundness proof, the completeness proof for K is easily adapted to other logics. To show that the T-rules are complete with respect to T-validity, for example, we merely need check that the model induced by any open branch on a fully developed T-tree is reflexive. It must be, because an open branch on a fully developed T-tree contains  $\omega R\omega$  for each world variable  $\omega$  on the branch.

**Exercise 4.7**

What do we need to check to show that the K4-rules are complete with respect to K4-validity?

**Exercise 4.8**

A Kripke model is *acyclical* if you can never return to the same world by following the accessibility relation. Show that if a sentence is true at some world in some Kripke model, then it is also true at some world in some acyclical Kripke model. (Hint: If  $A$  is true at some world in some Kripke model then  $\neg A$  is K-invalid.)

By the soundness theorem, there is a fully developed K-tree for  $\neg A$  with an open branch. Now consider the model induced by this branch.)

#### Exercise 4.9

The S5 tree rules from chapter 2 are sound and complete for S5-validity: all and only the S5-valid sentences can be proven. Are the rules sound for K-validity? Are they complete for K-validity?

## 4.4 Soundness and completeness for axiomatic calculi

Next, we are going to show that the axiomatic calculus for system K is sound and complete for K-validity. In the axiomatic calculus, a proof is a list of sentences each of which is either an instance of (Dual) or (K) or can be derived from earlier sentences on the list by application of (CPL) or (Nec). Expressed as a construction rule, (Nec) says that whenever a list contains a sentence  $A$  then one may append  $\Box A$ . (CPL) says that one may append any truth-functional consequence of sentences that are already on the list. (This is an acceptable rule because there is a simple mechanical test – the truth-table method – for checking whether a sentence is a truth-functional consequence of finitely many other sentences.)

Soundness is easy. We want to show that everything that is derivable from some instances of (Dual) and (K) by applications of (CPL) and (Nec) is K-valid. We show this by showing that (1) every instance of (Dual) and (K) is K-valid, and (2) every sentence that is derived from K-valid sentences by (CPL) or (Nec) is itself K-valid.

#### Theorem: Soundness of the axiomatic calculus for K

Any sentence that is provable in the axiomatic calculus for K is K-valid.

*Proof:* We first show that every instance of (Dual) and (K) is K-valid.

1. (Dual) is the schema  $\neg\Diamond A \leftrightarrow \Box\neg A$ . By clauses (b), (g), and (h) of definition 3.2, a sentence  $\neg\Diamond A$  is true at a world  $w$  in a Kripke model  $M$  iff  $\Box\neg A$  is true at  $w$  in  $M$ . It follows by clauses (f) and (e) that all instances of (Dual) are true at all worlds in all Kripke models.
2. (K) is the schema  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ . By clause (e) of definition 3.2, a sen-

tence  $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$  is false at a world  $w$  in a Kripke model only if  $\Box(A \rightarrow B)$  and  $\Box A$  are both true at  $w$  while  $B$  is false. By clause (g) of definition 3.2,  $\Box B$  is false at  $w$  only if  $B$  is false at some world  $v$  accessible from  $w$ . But if  $\Box(A \rightarrow B)$  and  $\Box A$  are both true at  $w$ , then  $A \rightarrow B$  and  $A$  are true at every world accessible from  $w$ , again by clause (g). And 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.

Next we show that (CPL) and (Nec) preserve K-validity.

1. By definition 3.2, the truth-functional operators have their standard truth-table meaning at every world in every Kripke model. It follows that all truth-functional consequences of sentences that are true at a world are themselves true at that world. In particular, if some sentences are true at every world in every Kripke model, then any truth-functional consequence of these sentences is also true at every world every Kripke model.
2. Let  $w$  be an arbitrary world in an arbitrary Kripke model. If  $A$  is true at every world in every Kripke model, then  $A$  is true at every world accessible from  $w$ , in which case  $\Box A$  is true at  $w$  by clause (g) of definition 3.2. So if  $A$  is K-valid, then  $\Box A$  is also K-valid.  $\square$

The soundness proof for K is easily extended to other modal systems. Since all instances of (Dual) and (K) are true at all worlds in all Kripke models, they are also true at all worlds in any more restricted class of Kripke models. The arguments for (CPL) and (Nec) also go through if we replace ‘every Kripke model’ by ‘every Kripke model of such-and-such type’. So if we want to show that, say, the axiomatic calculus for T is sound with respect to the concept of T-validity – that is, if we want to show that anything that is derivable from (Dual), (K), and (T) by (CPL) and (Nec) is true at all worlds in all reflexive Kripke models – all that is left to do is to show that every instance of the (T)-schema is true at all worlds in all reflexive Kripke model. (We’ve already shown this: see observation 3.2.)

#### Exercise 4.10

Outline the soundness proof for the axiomatic calculus for S4, whose axiom schemas are (Dual), (K), (T), and (4).

Let’s turn to completeness. We are going to show that every K-valid sentence is derivable from some instances of (Dual) and (K) by (CPL) and (Nec). As in section 4.3, we

argue by contraposition. We will show that any sentence that cannot be derived from (Dual) and (K) by (CPL) and (Nec) is not K-valid. To show that a sentence is not K-valid, we will give a countermodel – a Kripke model in which the sentence is false at some world. In fact, we will give the *same* countermodel for every sentence that isn't derivable in the calculus. You might think we need different countermodels for different sentences, but it turns out that there is a particular model in which every K-invalid sentence is false at some world. This model is called the *canonical model* for K.

In order to define the canonical model, let's introduce some shorthand terminology. We'll say that an  $\mathcal{L}_M$ -sentence is *K-provable* if it can be proved in the axiomatic calculus for K. A set of  $\mathcal{L}_M$ -sentences is *K-inconsistent* if it contains a finite number of sentences  $A_1, \dots, A_n$  such that  $\neg(A_1 \wedge \dots \wedge A_n)$  is K-provable. A set is *K-consistent* if it is not K-inconsistent.

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

A set of  $\mathcal{L}_M$ -sentences is called *maximal* if it contains either  $A$  or  $\neg A$  for every  $\mathcal{L}_M$ -sentence  $A$ . A set is *maximal K-consistent* if it is both maximal and K-consistent.

#### Exercise 4.11

Which, if any, of these sets are K-consistent? (a)  $\{p\}$ , (b)  $\{\neg p\}$ , (c) the set of all sentence letters, (d) the set of all  $\mathcal{L}_M$ -sentences.

Now here's the canonical model for K.

#### Definition 4.3

The **canonical model**  $M_K$  for K is the Kripke model  $\langle W, R, V \rangle$ , where

- $W$  is the set of all maximal K-consistent sets of  $\mathcal{L}_M$ -sentences,
- $wRv$  iff  $v$  contains every sentence  $A$  for which  $w$  contains  $\Box A$ ,
- for every sentence letter  $P$ ,  $V(P)$  is the set of all members of  $W$  that contain  $P$ .

The “worlds” in the canonical model are sets of  $\mathcal{L}_M$ -sentences. The interpretation function makes a sentence letter true at a world iff the letter is a member of the world. As we are going to see, this generalizes to arbitrary sentences:

- (1) A world  $w$  in  $M_K$  contains all and only the sentences that are true at  $w$  in  $M_K$ .

We will also prove the following:

- (2) If some sentence cannot be proved in the axiomatic calculus for  $K$ , then its negation is a member of some world in  $M_K$ .

Together, these two lemmas will establish completeness for the axiomatic calculus. Fact (2) tells us that if a sentence  $A$  isn't  $K$ -provable, then  $\neg A$  is a member of some world  $w$  in the canonical model  $M_K$ . By fact (1), we can infer that  $\neg A$  is true at  $w$  in  $M_K$ , which means that  $A$  is false at  $w$  in  $M_K$ . So any sentence that isn't  $K$ -provable isn't  $K$ -valid.

We are going to prove (2) first. We'll need the following observation.

**Observation 4.2:** If a set  $\Gamma$  is  $K$ -consistent, then for any sentence  $A$ , either  $\Gamma \cup \{A\}$  or  $\Gamma \cup \{\neg A\}$  is  $K$ -consistent.

( $\Gamma \cup \{A\}$ , called the *union* of  $\Gamma$  and  $\{A\}$ , is the smallest set that contains all members of  $\Gamma$  as well as  $A$ .)

*Proof:* Let  $\Gamma$  be any  $K$ -consistent set and  $A$  any sentence. Suppose for reductio that  $\Gamma \cup \{A\}$  and  $\Gamma \cup \{\neg A\}$  are both  $K$ -inconsistent.

That  $\Gamma \cup \{A\}$  is  $K$ -inconsistent means there are sentences  $A_1, \dots, A_n$  in  $\Gamma \cup \{A\}$  such that  $\neg(A_1 \wedge \dots \wedge A_n)$  is  $K$ -provable. Since  $\Gamma$  itself is  $K$ -consistent, one of the sentences  $A_1, \dots, A_n$  must be  $A$ . Let  $B$  be the conjunction of the other sentences in  $A_1, \dots, A_n$ , all of which are in  $\Gamma$ . So  $\neg(B \wedge A)$  is  $K$ -provable.

That  $\Gamma \cup \{\neg A\}$  is  $K$ -inconsistent means that there are sentences  $A_1, \dots, A_n$  in  $\Gamma \cup \{\neg A\}$  such that  $\neg(A_1 \wedge \dots \wedge A_n)$  is  $K$ -provable. As before, one of these sentences must be  $\neg A$ . Let  $C$  be the conjunction of the others, all of which are in  $\Gamma$ . So  $\neg(C \wedge \neg A)$  is  $K$ -provable.

If  $\neg(B \wedge A)$  and  $\neg(C \wedge \neg A)$  are both  $K$ -provable, then so is  $\neg(B \wedge C)$ , because it is a truth-functional consequence of  $\neg(B \wedge A)$  and  $\neg(C \wedge \neg A)$ . But  $B \wedge C$  is a conjunction of sentences from  $\Gamma$ . So  $\Gamma$  itself is  $K$ -inconsistent, contradicting our assumption.  $\square$

Now we can prove fact (2).

#### Lindenbaum's Lemma

Every  $K$ -consistent set is a subset of some maximal  $K$ -consistent set.

*Proof:* Let  $S_0$  be some K-consistent set of sentences. Let  $A_1, A_2, \dots$  be a list of all  $\mathcal{L}_M$ -sentences in some arbitrary order. For every number  $i \geq 0$ , define

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

This gives us an infinite list of sets  $S_0, S_1, S_2, \dots$ . Each set in the list is K-consistent:  $S_0$  is K-consistent by assumption. And if some set  $S_i$  in the list is K-consistent, then either  $S_i \cup \{A_i\}$  is K-consistent, in which case  $S_{i+1} = S_i \cup \{A_i\}$  is K-consistent, or  $S_i \cup \{A_i\}$  is not K-consistent, in which case  $S_{i+1} = S_i \cup \{\neg A_i\}$ , which is K-consistent by observation 4.2. So if any set in the list is K-consistent, then the next set in the list is also K-consistent. It follows that  $S_0, S_1, S_2, \dots$  are all K-consistent.

Now let  $S$  be the set of sentences that occur in at least one of the sets  $S_0, S_1, S_2, S_3, \dots$ . (That is, let  $S$  be the union of  $S_0, S_1, S_2, S_3, \dots$ .) Evidently,  $S_0$  is a subset of  $S$ . And  $S$  is maximal. Moreover,  $S$  is K-consistent. For if  $S$  were not K-consistent, then it would contain some sentences  $B_1, \dots, B_n$  such that  $\neg(B_1 \wedge \dots \wedge B_n)$  is K-provable. All of these sentences would have to occur somewhere on the list  $A_1, A_2, \dots$ . Let  $A_j$  be a sentence from  $A_1, A_2, \dots$  that occurs after all the  $B_1, \dots, B_n$ . If  $B_1, \dots, B_n$  are in  $S$ , they would have to be in  $S_j$  already, so  $S_j$  would be K-inconsistent. But we've seen that all of  $S_0, S_1, S_2, \dots$  are K-consistent.  $\square$

Notice that the proof of Lindenbaum's Lemma does not turn on any assumptions about the axiomatic calculus for K except that (CPL) is one of its rules. The lemma holds for every calculus with (CPL) as a (possibly derived) rule.

To prove fact (1), we need another observation, which relies on the presence of (K) and (Nec), besides (CPL).

**Observation 4.3:** If  $\Gamma$  is a maximal K-consistent set of sentences that does not contain  $\Box A$ , and  $\Gamma^-$  is the set of all sentences  $B$  for which  $\Box B$  is in  $\Gamma$ , then  $\Gamma^- \cup \{\neg A\}$  is K-consistent.

*Proof:* We show that if  $\Gamma^- \cup \{\neg A\}$  is not K-consistent, then neither is  $\Gamma$ . If  $\Gamma^- \cup \{\neg A\}$  is not K-consistent, then there are sentences  $B_1, \dots, B_n$  in  $\Gamma^-$  such that  $\neg(B_1 \wedge \dots \wedge B_n \wedge \neg A)$  is K-provable. And then  $(B_1 \wedge \dots \wedge B_n) \rightarrow A$  is K-provable, because it is a truth-functional consequence of  $\neg(B_1 \wedge \dots \wedge B_n \wedge \neg A)$ . By repeated application of (Nec), (K), and (CPL), one can derive  $(\Box B_1 \wedge \dots \wedge \Box B_n) \rightarrow \Box A$  from  $(B_1 \wedge \dots \wedge B_n) \rightarrow A$ . Another application of (CPL) yields  $\neg(\Box B_1 \wedge \dots \wedge \Box B_n \wedge \neg \Box A)$ . So  $\{\Box B_1, \dots, \Box B_n, \neg \Box A\}$

is K-inconsistent. But  $\Box B_1, \dots, \Box B_n$  are in  $\Gamma$ . And since  $\Box A$  is not in  $\Gamma$  and  $\Gamma$  is maximal,  $\neg \Box A$  is in  $\Gamma$ . So  $\{\Box B_1, \dots, \Box B_n, \neg \Box A\}$  is a subset of  $\Gamma$ . And so  $\Gamma$  is K-inconsistent.  $\square$

Here, then, is fact (1):

#### Canonical Model Lemma

For any world  $w$  in  $M_K$  and any sentence  $A$ ,  $A$  is in  $w$  iff  $M_K, w \models A$ .

*Proof:* The proof is by induction on complexity of  $A$ . We first show that the claim (that  $A$  is in  $w$  iff  $M_K, w \models A$ ) holds for sentence letters. Then we show that if the claim holds for the immediate parts of a complex sentence (this is our induction hypothesis), then the claim also holds for the sentence itself.

- Suppose  $A$  is a sentence letter. By definition 4.3,  $w \in V(A)$  iff  $A \in w$ . So by clause (a) of definition 3.2,  $M_K, w \models A$  iff  $A \in w$ . (' $\in$ ' means 'is a member of the set'.)
- Suppose  $A$  is a negation  $\neg B$ . By clause (b) of definition 3.2,  $M_K, w \models \neg B$  iff  $M_K, w \not\models B$ . By induction hypothesis,  $M_K, w \not\models B$  iff  $B \notin w$ . Since  $w$  is maximal K-consistent,  $B \notin w$  iff  $\neg B \in w$ . So  $M_K, w \models \neg B$  iff  $\neg B \in w$ .
- Suppose  $A$  is a conjunction  $B \wedge C$ . By clause (c) of definition 3.2,  $M_K, w \models B \wedge C$  iff  $M_K, w \models B$  and  $M_K, w \models C$ . By induction hypothesis,  $M_K, w \models B$  iff  $B \in w$ , and  $M_K, w \models C$  iff  $C \in w$ . Since  $w$  is maximal K-consistent,  $B$  and  $C$  are in  $w$  iff  $B \wedge C$  is in  $w$ . So  $M_K, w \models B \wedge C$  iff  $B \wedge C \in w$ .

The cases for the other truth-functional connectives are similar.

- Suppose  $A$  is a box sentence  $\Box B$ , and that  $\Box B \in w$ . By definition 4.3, it follows that  $B \in v$  for all  $v$  with  $wRv$ . By induction hypothesis, this means that  $M_K, v \models B$  for all  $v$  with  $wRv$ . And then  $M_K, w \models \Box B$ , by clause (g) of definition 3.2.

For the converse direction, suppose  $\Box B \notin w$ . Let  $\Gamma^-$  be the set of all sentences  $C$  for which  $\Box C \in w$ . By observation 4.3,  $\Gamma^- \cup \{\neg B\}$  is K-consistent. By definition 4.3 and Lindenbaum's Lemma, it follows that there is some  $v \in W$  such that  $wRv$  and  $\neg B \in v$ . Since  $v$  is K-consistent,  $B \notin v$ . By induction hypothesis, it follows that  $M_K, v \not\models B$ . And so  $M_K, w \not\models \Box B$ , by clause (g) of definition 3.2.

- Suppose  $A$  is a diamond sentence  $\Diamond B$ , and that  $\Diamond B \in w$ . By (Dual) and (CPL), any set that contains both  $\Diamond B$  and  $\Box \neg B$  is K-inconsistent. So  $\Box \neg B \notin w$ . By observation 4.3 and Lindenbaum's Lemma (as in the previous case), it follows that there is some

$v \in W$  such that  $wRv$  and  $B \in v$ . By induction hypothesis,  $M, v \models B$ . So  $M_K, w \models \Diamond B$ , by clause (h) of definition 3.2.

For the converse direction, suppose  $\Diamond B \notin w$ . Then  $\Box \neg B \in w$ , by (Dual), (CPL), and the fact that  $w$  is maximal K-consistent. By definition 4.3, it follows that  $\neg B \in v$  for all  $v$  with  $wRv$ . Since all such  $v$  are maximal K-consistent, none of them contain  $B$ . By induction hypothesis,  $B$  is not true at any of them. By clause (h) of definition 3.2, it follows that  $M_K, w \not\models \Diamond B$ .  $\square$

The completeness of the axiomatic calculus for K follows immediately from the previous two lemmas, as foreshadowed above:

**Theorem: Completeness of the axiomatic calculus for K**

If  $A$  is K-valid, then  $A$  is provable in the axiomatic calculus for K.

*Proof:* We show that if a sentence is not K-provable then it is not K-valid. Suppose  $A$  is not K-provable. Then  $\{\neg A\}$  is K-consistent. It follows by Lindenbaum's Lemma that  $\{\neg A\}$  is included in some maximal K-consistent set  $S$ . By definition 4.3, that set is a world in  $M_K$ . Since  $\neg A$  is in  $S$ , it follows from the Canonical Model Lemma that  $M_K, S \models \neg A$ . So  $M_K, S \not\models A$ . So  $A$  is not true at all worlds in all Kripke models.  $\square$

Done!

Once again, the proof is easily adjusted to many axiomatic calculi for logics stronger than K. All we have assumed about the K-calculus is that it contains (Dual), (K), (Nec), and (CPL). So if we're interested in, say, whether the axiomatic calculus for T is complete, we can simply replace 'K-consistent' by 'T-consistent' throughout the proof, and almost everything goes through as before. We only have to add a small step at the end.

By adapting the argument for K, we can show that if a sentence  $A$  is not T-provable then  $A$  is false at some world in the canonical model for T. This shows that  $A$  is not K-valid. But we want to show that  $A$  is not T-valid – meaning that  $A$  is not true at all worlds in all reflexive Kripke models. To complete the proof, we need to show that the canonical model  $M_T$  for T is reflexive.

This isn't hard. Given how accessibility in canonical models is defined, 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  $M_T$  are maximal T-consistent sets of sentences, and every such set contains every instance of the (T) schema  $\Box A \rightarrow A$ , there is no world in  $M_T$  that contains  $\Box A$  but not  $A$ . So every world in  $M_T$  has access to itself.

In general, to show that a calculus that extends the K-calculus by further axiom schemas

is complete, we only need to show that the canonical model for the calculus satisfies the frame conditions that correspond to the added axiom schemas. This is usually the case. But not always. Sometimes, an axiomatic calculus is sound and complete with respect to some class of Kripke models, but the canonical model of the calculus is not a member of that class. (An example is the calculus for the system GL, which I will describe at the very end of this chapter.) Completeness must then be established by some other means.

#### Exercise 4.12

Outline the completeness proof for the axiomatic calculus for S5.

#### Exercise 4.13

The set of all  $\mathcal{L}_M$ -sentences is a system of modal logic. Let's call this system  $X$  (for "explosion"). (a) Describe a sound and complete proof method for  $X$ . (b) Explain why  $X$  does not have a canonical model.

## 4.5 Loose ends

You will remember from observation 1.1 in chapter 1 that claims about entailment can be converted into claims about validity.  $A$  entails  $B$  iff  $A \rightarrow B$  is valid;  $A_1$  and  $A_2$  together entail  $B$  iff  $A_1 \rightarrow (A_2 \rightarrow B)$  – equivalently,  $(A_1 \wedge A_2) \rightarrow B$  – is valid; and so on. But what if there are infinitely many premises  $A_1, A_2, A_3, \dots$ ? Sentences of  $\mathcal{L}_M$  are always finite, so we can't convert the claim that  $A_1, A_2, A_3, \dots$  entail  $B$  into a claim that some  $\mathcal{L}_M$ -sentence is valid.

We also can't use the tree method or the axiomatic method to directly show that a conclusion follows from infinitely many premises. A proof in either method is a finite object that can only invoke finitely many sentences.

As it turns out, this is not a serious limitation. In many logics – including classical propositional and predicate logic and all the modal logics we have so far encountered – a sentence is entailed by infinitely many premises only if it is entailed by a finite subset of these premises. Logics with this property are called **compact**.

Let's show that  $K$  is compact. To this end, I'll say that a sentence  $B$  is *K-derivable* from a (possibly infinite) set of sentences  $\Gamma$  if there are finitely many members  $A_1, \dots, A_n$  of  $\Gamma$  for which  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$  is provable in the axiomatic calculus for  $K$ . Now we first show that whenever  $\Gamma \models_K B$  then  $B$  is  $K$ -derivable from  $\Gamma$ . This is called *strong*

*completeness* because it is stronger than the (“weak”) kind of completeness that we have established in the previous section.

**Theorem: Strong completeness of the axiomatic calculus for K**

Whenever  $\Gamma \models_K B$  then  $B$  is K-derivable from  $\Gamma$ .

*Proof:* Suppose  $B$  is not K-derivable from  $\Gamma$ . Then there are no  $A_1, \dots, A_n$  in  $\Gamma$  such that  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$  is K-provable. This means that  $\Gamma \cup \{\neg B\}$  is K-consistent. By Lindenbaum’s Lemma, it follows that  $\Gamma \cup \{\neg B\}$  is included in some maximal K-consistent set and thereby in some world in the canonical model  $M_K$  for K. (Lindenbaum’s lemma says that every K-consistent set of  $\mathcal{L}_M$ -sentences, even if it is infinite, is included in a maximal K-consistent set.) By the Canonical Model Lemma,  $M_K, w \models_K A$  for all  $A$  in  $\Gamma$ , and  $M_K, w \not\models_K B$ . Thus  $\Gamma \not\models_K B$ .  $\square$

**Theorem: Compactness of K**

If a sentence  $B$  is K-entailed by some sentences  $\Gamma$ , then  $B$  is K-entailed by a finite subset of  $\Gamma$ .

*Proof:* Suppose  $\Gamma \models_K B$ . By strong completeness, it follows that there are finitely many sentences  $A_1, \dots, A_n$  in  $\Gamma$  for which  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$  is K-provable. By the soundness of the K-calculus,  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$  is valid. So  $A_1, \dots, A_n \models_K B$ , by observation 1.1.  $\square$

Compactness is surprising. It is easy to think of cases in which a conclusion is entailed by infinitely many premises, but not by any finite subset of these premises. For example, suppose I like the number 0, I like the number 1, I like the number 2, and so on, for all natural numbers 0,1,2,3,... Together, these assumptions entail that I like every natural number. But no finite subset of the assumptions has this consequence.

**Exercise 4.14**

A set of sentences  $\Gamma$  is called *K-satisfiable* if there is a world in some Kripke model at which all members of  $\Gamma$  are true. Show that an infinite set of sentences  $\Gamma$  is K-satisfiable iff every finite subset of  $\Gamma$  is K-satisfiable.

To conclude this chapter, I want to take a quick look at the logic of mathematical

provability.

Our proofs of soundness, completeness, compactness, etc. were informal. We have not translated the relevant claims into a formal language, nor have we used a formal method of proof. In principle, however, this can be done. All our proofs could be formalized in an axiomatic calculus for predicate logic with a few additional axioms about sets. A well-known calculus of that kind is ZFC (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 practically everything that can be proved in any branch of maths.

An interesting feature of ZFC is that it can not only prove facts about what's provable in simpler axiomatic calculi; it can also prove facts about what's provable in ZFC itself. For example, one can prove in ZFC that one can prove in ZFC that  $2+2=4$ .

This gives us an interesting application of modal logic. Let's read the box as 'it is mathematically provable that', which we understand as provability in ZFC. One can easily show (in ZFC) that this operator has all the properties of the box in the basic logic K. For example, all instances of the (K)-schema are provable in ZFC. (The language of ZFC doesn't have a box symbol. But one can encode the (K)-schema into a schema of ZFC, given the present reading of the box, and all instances of that schema are ZFC-provable.)

So the logic of mathematical provability is at least as strong as K. In fact, it is stronger. One can prove in ZFC that whenever a sentence is ZFC-provable then it is ZFC-provable that the sentence is ZFC-provable. This gives us the (4)-schema  $\Box A \rightarrow \Box \Box A$ .

You might expect that we also have the (T)-schema  $\Box A \rightarrow A$  or the (D)-schema  $\Box A \rightarrow \Diamond A$ . The latter says that if something is provable then its negation isn't provable (since  $\Diamond A$  means  $\neg \Box \neg A$ ). And surely ZFC can't prove both a sentence and its negation – which would make ZFC inconsistent. I say 'surely', but can we prove (in ZFC) that ZFC is consistent? The answer is no. More precisely, one can prove that if one can prove that ZFC is consistent then ZFC is *inconsistent*. This bizarre fact is a consequence of *Gödel's second incompleteness theorem*, established by Kurt Gödel in 1931. It is reflected by the following schema (named after Gödel and Martin Löb), all whose instances are provable in ZFC:

$$(GL) \quad \Box(\Box A \rightarrow A) \rightarrow \Box A$$

The system GL, which is axiomatized by (K), (GL), (Nec), and (CPL), completely captures what ZFC can prove about provability in ZFC. (Schema (4) isn't needed as a separate axiom schema because it can be derived.)

**Exercise 4.15**

Suppose ZFC can prove its own consistency, so that there is a proof of  $\neg\Box(p \wedge \neg p)$ . Explain how this proof could be extended to a proof of  $\Box(p \wedge \neg p)$ . You need each of (GL), (Nec), and (CPL).

## 5 Epistemic Logic

### 5.1 Epistemic accessibility

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

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

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

#### Exercise 5.1

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

We will interpret the box and the diamond in terms of epistemic accessibility. In this context, the box is usually written ‘K’. For once, this doesn’t stand for Kripke but for knowledge. I will use ‘M’ (‘might’) for the diamond. So  $KA$  means that  $A$  is true at all epistemically accessible worlds, while  $MA$  means that  $A$  is true at some epistemically accessible world. If we want to clarify which agent we have in mind, we can add a subscript:  $M_b A$  might say that  $A$  is epistemically possible for Bob.

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

To see why, note that if some propositions are true at a world, then anything that logically follows from these propositions is also true at that world. For example, if  $p \rightarrow q$  and  $p$  are both true at  $w$ , then so is  $q$  (by definition 3.2). As a consequence, if  $p \rightarrow q$  and  $p$  are true at all epistemically accessible worlds (for some agent), then  $q$  is also true at all these worlds.  $K(p \rightarrow q)$  and  $Kp$  together entail  $Kq$ . More generally, the  $K$  operator is **closed under logical consequence**, meaning that if  $B$  logically follows from  $A_1, \dots, A_n$ , and  $KA_1, \dots, KA_n$ , then  $KB$ .

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

### Exercise 5.2

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

- (a) Alice knows that it is either raining or snowing.
- (b) Either Alice knows that it is raining or that it is snowing.
- (c) Alice knows whether it is raining.
- (d) You know that you’re guilty if you don’t know that you’re innocent.

## 5.2 The logic of knowledge

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

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

$$(T) \quad KA \rightarrow A$$

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

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

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

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

My second argument starts with a scenario in which someone has misleading evidence that some proposition  $p$  is false. This is easily conceivable. In that scenario,  $p$  is true but the agent believes  $\neg p$ . Often, when we believe something, we also believe that we know it. Let's assume that our agent believes that they know  $\neg p$ . Let's also assume that their beliefs are consistent, so they don't believe that they *don't* know  $\neg p$ . Since they don't believe this proposition (that they don't know  $\neg p$ ) they don't know it either: they don't know that they don't know  $\neg p$ . So we have a scenario in which  $p$  is true but  $K\neg K\neg p$  false.

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

$$(B) \quad A \rightarrow KMA.$$

This means that all instances of (B) are valid on a frame iff the frame is symmetric. If the epistemic accessibility relation were symmetric, then all instances of (B) would be valid.

But I've just described a scenario in which an instance of (B) is false. So the epistemic accessibility relation isn't symmetric.

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

$$(4) \quad KA \rightarrow KKA$$

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

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

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

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

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

We can formalize this concept in modal logic. Let  $N$  mean that conditions are normal (whatever exactly this means), and let  $\Box$  be a non-epistemic operator that formalizes ‘at all worlds’.  $\Box(N \rightarrow A)$  then means that  $A$  is true at all world at which conditions are normal. According to the definition I just gave, a belief state  $s$  indicates  $p$  iff

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

The state  $s$  indicates that  $s$  indicates  $p$  iff

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

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

### Exercise 5.3

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

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

$$(5) \quad M A \rightarrow K M A.$$

We know that the (5)-schema corresponds to euclidity. This gives us a quick argument against the schema. As you showed in exercise 3.11, reflexivity and euclidity together entail symmetry. The epistemic accessibility relation is reflexive. If it were euclidean, it would be symmetric. But I've argued that it isn't symmetric. So the logic of knowledge doesn't validate (5).

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

Here it is important to not be misled by a curiosity of ordinary language. When we say that someone doesn't know  $p$ , this seems to imply that  $p$  is true. If I told you that my neighbour doesn't know that I have a pet aardvark, you could reasonably infer that I have a pet aardvark. You might therefore be tempted to regard all instances of the following schema as valid:

$$(NT) \quad \neg K A \rightarrow A$$

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

#### Exercise 5.4

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

#### Exercise 5.5

Let's say that an agent is *ignorant of* a proposition if they don't know the proposition and the proposition is true. (In English, saying that someone doesn't know a proposition normally conveys that they are ignorant of the proposition, in this sense.) Show that if the logic of knowledge is at least as strong as K, then ignorance of  $A$  entails ignorance of ignorance of  $A$ .

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

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

$$(0.4) \quad (\neg KA \wedge A) \rightarrow K\neg KA$$

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

### Exercise 5.6

Explain why Gettier cases cast doubt on (0.4).

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

$$(G) \quad M K A \rightarrow K M A$$

The resulting logic is called S4.2; it is weaker than S4.4 but stronger than S4. We will meet an argument in favour of (G) in section 5.4.

### Exercise 5.7

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

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

## 5.3 Multiple Agents

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

We can easily expand the language  $\mathcal{L}_M$  to a **multi-modal language** by introducing a whole series of box operators  $K_1, K_2, K_3, \dots$  with their duals  $M_1, M_2, M_3, \dots$ . This multi-modal language is interpreted in multi-modal Kripke models.

### Definition 5.1

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

- a non-empty set  $W$ ,
- a set of binary relation  $R_1, R_2, R_3, \dots$  on  $W$ , and

- a function  $V$  that assigns to each sentence letter a subset of  $W$ .

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

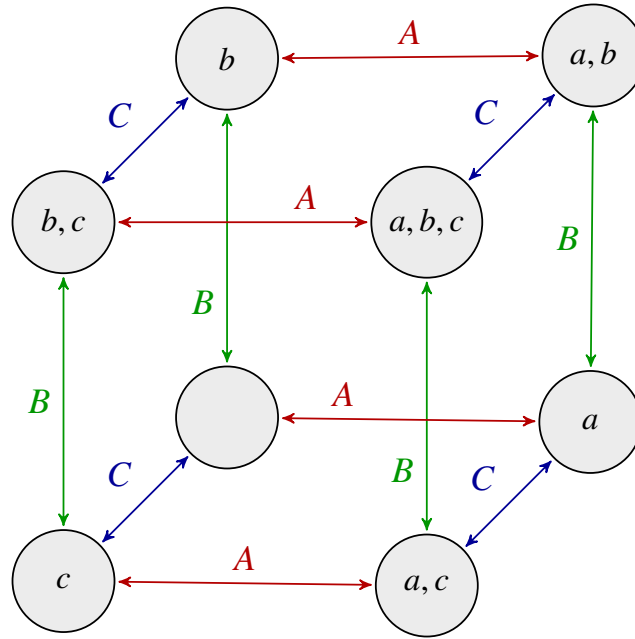
The definition of truth at a world in a Kripke model (definition 3.2) is easily extended to multi-modal Kripke models. Instead of clauses (g) and (h), we have the following conditions, for each pair of a modal operator ( $K_i$  or  $M_i$ ) and the corresponding accessibility relation  $R_i$ :

$$\begin{aligned} M, w \models K_i A & \quad \text{iff } M, v \models A \text{ for all } v \text{ in } W \text{ such that } wR_i v. \\ M, w \models M_i A & \quad \text{iff } M, v \models A \text{ for some } v \text{ in } W \text{ such that } wR_i v. \end{aligned}$$

For an application of this machinery, let's look at the *Muddy Children* puzzle.

Three (intelligent) children have been playing outside. They can't see or feel if their own face is muddy, but they can see who of the others have mud on their face. As they come inside, mother tells them: 'At least one of you has mud on their face'. She then asks, 'Do you know if you have mud on your face?'. All three children say that they don't know. Mother asks again, 'Do you know if you have mud on your face?'. This time, two children say that they know. How many children have mud on their face? What happens if the mother asks her question a third time?

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

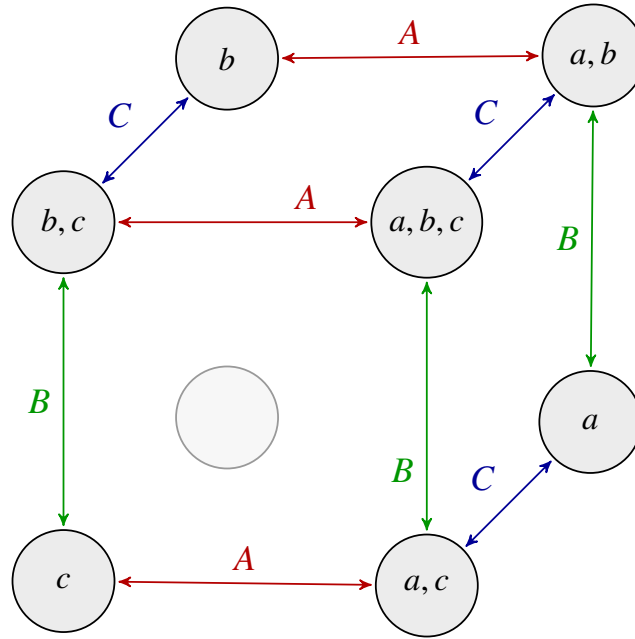


Since we have three epistemic agents, we have three accessibility relations, one for Alice (drawn in red), one for Bob (green), and one for Carol (blue). To remove clutter, I have left out the  $(3 \times 8)$  arrows leading from each world to itself, but we should keep in mind that every world is also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

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

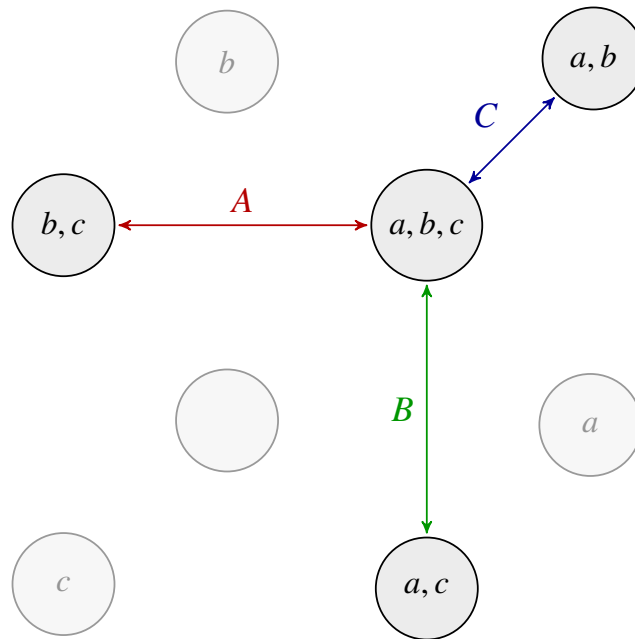
Notice how the fact that every child can see the others is reflected in the diagram. For example, at the top left world, where only Bob is muddy, Alice sees that Bob is muddy and that Carol is clean; the only epistemic possibilities for Alice at that world are the two worlds at the top: the  $b$  world itself and the  $a, b$  world to the right. In general, the only accessible worlds for a given child at a given world  $w$  are worlds at which the other children's state of muddiness is the same as at  $w$ .

What changes through the mother's first announcement, 'At least one of you has mud on their face'? The announcement tells *us* that we're not in the world where  $a, b$ , and  $c$  are all false. More importantly, it allows *each child* to rule out this world (since they all hear and accept the announcement).



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

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



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

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

#### Exercise 5.8

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

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

‘My birthday is one of these’, she says. Then she announces that she will whisper the month of her birthday in Albert’s ear and the day in Bernard’s. After the

whispering, she asks Albert if he knows her birthday. Albert says, ‘no, but I know that Bernard doesn’t know either’. To which Bernard responds: ‘Right. I didn’t know until now, but now I know’. Albert: ‘Now I know too!’ Draw a multi-modal Kripke model for each stage of the conversation. When is Cheryl’s birthday?

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

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

$$K_1 K_2 A \rightarrow K_1 A.$$

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

A common assumption in epistemic logic is that there are no genuinely new interaction principles for the knowledge of multiple agents – no principles that don’t already follow from the logic of individual knowledge. The above principle, for example, is entailed by the assumption that the (T)-schema holds for  $K_2$ . Think of the relevant Kripke models. Suppose, as  $K_1 K_2 A$  asserts, that  $A$  holds at each world that is  $R_2$ -accessible from any  $R_1$ -accessible world. If the (T)-schema holds for  $K_2$ , then every world is  $R_2$ -accessible from itself. In particular, then, any  $R_1$ -accessible world is  $R_2$ -accessible from itself. It follows that  $A$  holds at every  $R_1$ -accessible world. So  $K_1 A$  is true.

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

Here is a tree proof of the schema  $K_1 K_2 A \rightarrow K_1 A$ , assuming that  $R_2$  is reflexive.

1.	$\neg(K_1 K_2 A \rightarrow K_1 A)$	(w)	(Ass.)
2.	$K_1 K_2 A$	(w)	(1)
3.	$\neg K_1 A$	(w)	(1)
4.	$wR_1 v$		(3)
5.	$\neg A$	(v)	(3)
6.	$K_2 A$	(v)	(2,4)
7.	$vR_2 v$		(Refl.)
8.	$A$	(v)	(6,7)
	x		

#### Exercise 5.9

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

- (a)  $M_1 K_2 p \rightarrow M_1 p$
- (b)  $M_1 K_2 p \rightarrow M_2 M_1 p$
- (c)  $M_1 K_2 p \rightarrow M_2 K_1 p$
- (d)  $K_1 K_2 p \rightarrow K_2 K_1 p$

We can also define new modal operators for groups of agents. A proposition is said to be **mutually known** in a group  $G$  if it is known by every member of the group. Let  $E_G$  be an operator for mutual knowledge. Clearly,  $E_G A$  can be defined as  $K_1 A \wedge K_2 A \wedge \dots \wedge K_n A$ , where  $K_1, K_2, \dots, K_n$  are the knowledge operators for the members of the group. So we can't say anything new with the help of  $E_G$  (at least for finite groups). But it can be instructive to see how  $E_G$  behaves depending on the behaviour of the underlying operators  $K_1, K_2$ , etc. For example, if each individual knowledge operator validates the (T)-schema, then so does  $E_G$ ; but if each  $K_i$  validates (4), it does not follow that  $E_G$  validates (4). For a counterexample, consider a group of two agents; both know  $p$ , and both know of themselves that they know  $p$ , but agent 1 does not know that agent 2 knows  $p$ . Then  $E_G p$  but  $\neg E_G E_G p$ .

#### Exercise 5.10

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

A more interesting concept that has proved useful in many areas is that of common knowledge. A proposition is **commonly known** in a group if everyone knows it, everyone knows that everyone knows it, everyone knows that everyone knows that everyone knows it, and so on forever. Let's use  $C_G$  as an operator for common knowledge.  $C_G$  is not definable in terms of  $K_1, \dots, K_n$ . Still, we can define it semantically in terms of the accessibility relations for the individual agents:  $C_G A$  is true at a world  $w$  iff  $A$  is true at all worlds that are reachable from  $w$  by some finite sequence of steps following the agents' accessibility relations.

It is easy to see that common knowledge validates (all instances of) (4). It validates (T) whenever individual knowledge validates (T). So the logic of common knowledge is at least S4. The complete logic of common knowledge also contain some non-trivial interaction principles, which are easiest to state in terms of  $E_G$ :

$$\begin{aligned} \text{(CK1)} \quad & C_G A \leftrightarrow (A \wedge E_G C_G A) \\ \text{(CK2)} \quad & (A \wedge C_G(A \rightarrow E_G A)) \rightarrow C_G A \end{aligned}$$

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

## 5.4 Knowledge, belief, and other modalities

Issues in the logic of knowledge can sometimes be clarified by looking at the connections between knowledge and belief. To formalise these connections, let's introduce a new operator  $B$  for belief – or rather, for *implicit belief*, since  $B$ , like  $K$ , will be closed under logical consequence.

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

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

The logic of  $B$  is different from the logic of  $K$ , if only because beliefs can be false. So

we will not regard all instances of

$$(T) \quad B A \rightarrow A$$

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

$$(D) \quad B A \rightarrow \neg B \neg A.$$

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

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

$$(4) \quad B A \rightarrow B B A$$

$$(5) \quad \neg B A \rightarrow B \neg B A$$

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

#### Exercise 5.11

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

#### Exercise 5.12

Show (in any way you like) that  $B(BA \rightarrow A)$  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 with both the  $K$  operator and the  $B$  operator. Models for this language will have two accessibility relations  $R_e$  and  $R_d$ . The first represents epistemic accessibility and is used for the interpretation of  $K$ , the second represents doxastic accessibility and is used to interpret  $B$ .

The power of combined logics for (implicit) knowledge and belief lies in the interaction principles that might link the two concepts. Here is a list of popular principles that

don't follow from the individual logics of knowledge and belief.

- (KB)  $KA \rightarrow BA$
- (PI)  $BA \rightarrow KBA$
- (NI)  $\neg BA \rightarrow K\neg BA$
- (SB)  $BA \rightarrow BKA$

(KB) assumes that knowledge implies belief. (PI) and (NI) strengthen the introspection principles for belief. They assume that a state of belief or disbelief is always known to the agent. (SB) assumes that if an agent believes something then they also believe that they know it. This is sometimes said to reflect a conception of “strong belief”, on which belief is incompatible with doubt. If you believe  $p$  in the sense that you have no doubt that  $p$ , then you plausibly believe that you know  $p$ .

These interaction principles, together with the (D)-schema for belief, imply that an agent believes a proposition just in case they don't know that they don't know it:

$$(BMK) \quad BA \leftrightarrow MKA$$

Somewhat surprisingly, then, we could define belief in terms of knowledge.

Here is how we can get from  $BA$  to  $MKA$ .

1. Suppose  $BA$ .
2. By (SB), it follows that  $BKA$ .
3. By (D), it follows that  $\neg B\neg KA$ .
4. By (KB), it follows that  $\neg K\neg KA$ , and so that  $MKA$ .

To show that  $MKA$  entails  $BA$ , I'll show that  $\neg BA$  entails  $\neg MKA$ .

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

Given the equivalence between  $BA$  and  $MKA$ , the (D)-schema for belief

$$BA \rightarrow \neg B\neg A$$

is equivalent to

$$MKA \rightarrow \neg MK\neg A$$

which in turn is equivalent to

$$MKA \rightarrow KMA.$$

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

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

#### Exercise 5.13

Show that the interaction principles entail principles (4) and (5) for belief:  
 $BA \rightarrow BBA$  and  $\neg B\neg A \rightarrow B\neg B\neg A$ .

#### Exercise 5.14

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

#### Exercise 5.15

There seems to be no natural expression in English for the dual of belief. A common way to express that someone does not believe not  $p$  is to say that they believe that it might be that  $p$ , which has the surface form  $\Box\Diamond p$ . Can you explain why this might be an adequate way of expressing  $\Diamond p$ ?

It can also be instructive to combine epistemic with non-epistemic operators. Philosophers have often been interested not just in what we *do* know, but also in what we *can* know. Various skeptical arguments, for example, suggest that we *cannot know* that we have hands. For another example, the “verificationist” movement in the early 20th century assumed that a sentence is meaningful only if its truth-value can in principle

be settled by mathematical proof or empirical investigation. This would imply that a sentence is meaningful only if *it is possible to know* that it is true.

We can formalize claims like these in a multi-modal language with a knowledge operator  $K$  and a diamond  $\Diamond$  for the relevant kind of circumstantial possibility. The verificationist hypothesis that every truth is in principle knowable is then expressed by the following interaction principle:

$$(\text{Knowability}) \quad A \rightarrow \Diamond K A$$

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

1. Let  $p$  be any unknown truth. (Nobody thinks all truths are actually known.)
2. So we have  $p \wedge \neg K p$ .
3. In any logic that extends the minimal system  $K$ ,  $K(p \wedge \neg K p)$  entails  $K p \wedge K \neg K p$ .
4. By the (T)-schema for knowledge,  $K \neg K p$  entails  $\neg K p$ .
5. So  $K(p \wedge \neg K p)$  entails both  $K p$  and  $\neg K p$ .
6. So the hypothesis  $K(p \wedge \neg K p)$  is inconsistent.
7. So  $\neg \Diamond K(p \wedge \neg K p)$ .
8. Lines 2 and 7 together provide a counterexample to the Knowability principle.

#### Exercise 5.16

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

## 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’). If we read  $q$  as stating that you cook dinner, we might use  $Oq$  to express that you are obligated to cook dinner.

We assume that obligation and permission are duals. You are not obligated to cook dinner iff you are permitted to not cook dinner; you are not permitted to cook dinner iff you are obligated to not cook dinner.

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 norms of all kinds. We do not have to settle whether  $O$  expresses legal obligation or moral obligation or some other kind of obligation. It is important, however, that we don’t equivocate. If the law requires  $q$  and morality  $\neg q$ , we should not formalize this as  $Oq \wedge O\neg q$ . It would be better to use a multi-modal language with different operators for legal and moral obligation.

Obligations and permissions often 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  $O_1q \wedge \neg O_2q$ . But what does the sentence letter  $q$  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,  $O$  and  $P$  are sentence operators. Unless we want to say that verb phrases in English (like ‘cook dinner’) should be translated into sentences of  $\mathcal{L}_M$  – which is possible, but non-standard – we have to transform the acts that appear to be the objects of obligation and permission into propositions.

Consider sentence (1), which is arguably equivalent to (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  $O_1 q$ , where  $q$  translates ‘you cook dinner’, and  $O_1$  corresponds to ‘you ought to see to it that’.

The subject (you) is mentioned twice in (2). A common assumption in deontic logic is that we can drop the agent subscripts from 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. Without further assumptions, this does not impose any obligations on anyone.

There are reasons to question the equivalence between agent-relative ‘ought’ statements like (2) and impersonal ‘ought’ statements like (3). Suppose Amy has promised to play with Betty. Then 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  $O_a \text{ STIT } p$ , where STIT formalizes ‘sees to it that’. This allows us to distinguish between the following three possibilities.

- |                            |   |
|----------------------------|---|
| $O_a \text{ STIT } \neg p$ | Amy ought to see to it that she doesn’t play with Betty.                  |
| $O_a \neg \text{ STIT } p$ | Amy ought to not see to it that she plays with Betty.                     |
| $\neg O_a \text{ STIT } p$ | 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 without agent subscripts. This approach is sufficient for many applications, but its limitations should be kept in mind.

**Exercise 6.1**

Translate the following sentences into the standard language of deontic logic (without STIT or agent subscripts).

- (a) You must not go into the garden.
- (b) You may not go into the garden.
- (c) Jones ought to help his neighbours.
- (d) If Jones is going to help his neighbours, then he ought to tell them he's coming.
- (e) If Jones isn't going to help his neighbours, then he ought to not tell them he's coming.

## 6.2 Standard deontic logic

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. Let's call a world *ideal* (relative to some norms) if everything that happens at the world conforms to the norms. In an ideal world, everyone does what they ought to do.

How do the ideal worlds relate to permission and obligation? For a start, everything that happens at an ideal world is plausibly permitted, for we know that it conforms to the norms. The converse is plausible as well: whenever something is permitted then it happens at some ideal world. For suppose something doesn't happen at any ideal world. Then the event entails the violation of some norm: it is incompatible with the satisfaction of all norms. And then it can't be permitted.

We have a simple possible-worlds analysis of permission:

$A$  is permitted (relative to some norms) iff  $A$  is the case at some world that is ideal (relative to these norms).

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

$A$  is obligatory (relative to some norms) iff  $A$  is the case at all worlds that are ideal (relative to these norms).

These analyses resemble the simple possible-worlds analysis from chapter 2, where we assumed that  $A$  is possible iff it is the case at some world, and necessary iff it is the case at all worlds. The difference is that we now quantify only over ideal worlds.

We can capture this restriction with the help of Kripke models. In Kripke semantics,  $\Diamond A$  is true at a world  $w$  iff  $A$  is true at some world that is accessible from  $w$ . Let's assume that a world is accessible, from any world, iff it is ideal. Then Kripke semantics implies that  $\Diamond A$  is true at  $w$  iff  $A$  is true at some ideal world. That's what we want.

The accessibility relation I've just defined is a little unusual: whether it holds between  $w$  and  $v$  does not at all depend on  $w$ . We have

$$wRv \text{ iff } v \text{ is ideal.}$$

But that's OK. The definition of a Kripke model allows for such degenerate relations.

Let's investigate the formal properties of our degenerate accessibility relation. Is it, say, reflexive? Transitive? Symmetric? Euclidean?

Transitivity says that if  $wRv$  and  $vRu$  then  $wRu$ . Now  $wRv$  means that  $v$  is ideal. And  $vRu$  means that  $u$  is ideal.  $wRu$  also means that  $u$  is ideal. So transitivity requires that if  $v$  is ideal and  $u$  is ideal then  $u$  is ideal. That's obviously true. So our accessibility relation is transitive. The same reasoning shows that it is euclidean. Our possible-worlds analysis therefore validates the (intuitively somewhat elusive) schemas (4) and (5).

$$(4) \quad OA \rightarrow OOA$$

$$(5) \quad PA \rightarrow OPA$$

What about reflexivity? The hypothesis that every world has access to itself would mean that every world is ideal. When we reason about permission and obligation, we normally don't take for granted that everyone does what they ought to do. We allow for the logical possibility that norms can be violated. So we don't assume that every world is ideal. Equivalently, we don't regard the (T)-schema

$$(T) \quad OA \rightarrow A$$

as valid.

### Exercise 6.2

Is the accessibility relation (as defined above) symmetric?

We might, however, impose the weaker condition of seriality. This would validate the

(D)-schema

$$(D) \quad OA \rightarrow PA.$$

Intuitively, (D) says that the norms are consistent: if you're obligated to do  $A$ , then you are not obligated to do not- $A$ . (Remember that  $PA$  is equivalent to  $\neg O\neg A$ .) Semantically, (D) corresponds to the assumption that there is at least one world at which all the norms are satisfied. If there were no such world, all sentences of the form  $OA$  would come out true, and all sentences of the form  $PA$  false. Everything would be obligatory, but nothing allowed. It is hard to make sense of such a scenario. If we 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. This 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.

Another weakening of reflexivity is "shift reflexivity".  $R$  is *shift reflexive* if  $wRv$  implies  $vRv$ : every world that can be seen can see itself. Shift reflexivity corresponds to the following schema (U) (for "utopia")

$$(U) \quad O(OA \rightarrow A)$$

In words: it ought to be the case that whatever ought to be the case is the case. Shift reflexivity is entailed by euclidity, so our logic validates (U).

### Exercise 6.3

Explain why euclidity entails shift reflexivity.

We could look at further properties of the accessibility relation, but we wouldn't find any plausible candidates that are not entailed by seriality, transitivity, and euclidity. The complete logic of obligation and permission, assuming the above possible-worlds analysis, is plausibly KD45.

We might, however, reconsider our analysis. We've assumed that there is a fixed set of norms that divides the worlds into "ideal" ones, where all the norms are respected, and "non-ideal" ones, where some norms are violated. We might call this an **absolutist** conception of norms. A **relativist** conception, by contrast, would allow that the norms may vary from world to world.

Suppose, for example, that we want to reason about what is required by the traffic laws. The traffic laws evidently vary from world to world. Consider a world at which cyclists are required to wear top hats. Norman is cycling in this world, without a top hat. Is he violating the traffic laws? He is violating the laws of *his* world, but not the laws of *our* world. On an absolutist approach, we interpret O and P as always referring to the laws of our world, no matter what world is under consideration:  $Op$  is true at  $w$  iff  $p$  is required by the laws of our world. On a relativist approach, we instead assume that  $Op$  is true at  $w$  iff  $p$  is required by the laws *at*  $w$ .

On the relativist approach, a world  $v$  is accessible from a world  $w$  iff everything that happens at  $v$  conforms to the (relevant) norms at  $w$ .

Transitivity and euclidity now become implausible. Let  $w$  be a world in which the only relevant norm is that one must drive on the left. Let  $v$  be a world in which everyone drives on the left, but the law allows driving on either side. Let  $u$  be a world in which some people drive on the right.  $v$  is accessible from  $w$  and  $u$  from  $v$ , but  $u$  is not accessible from  $w$ . We don't have transitivity.

#### Exercise 6.4

Show that the deontic accessibility relation is neither euclidean nor shift-reflexive, on the relativist approach.

As before, we probably don't want to assume reflexivity, but we might want to assume seriality, which now means that there is no world at which the norms make inconsistent demands.

The relativist conception seems to be more common in deontic logic. So-called **standard deontic logic** assumes only that the accessibility relation is serial, making the system D the complete logic of obligation and permission.

The absolutist logic KD45 and the relativist logic D can be shown to disagree only about sentences in which a deontic operator occurs in the scope of another deontic operator. Any sentence that does not contain an O or P operator embedded under another O or P operator is D-valid iff it is KD45-valid.

### Exercise 6.5

Use the tree method to check which of the following sentences are D-valid and which are KD45-valid.

- (a)  $P(p \vee q) \rightarrow (Pp \wedge Pq)$
- (b)  $OPp \rightarrow Pp$
- (c)  $\neg P(p \vee q) \rightarrow (P\neg p \vee P\neg q)$
- (d)  $OPp \vee POp$

### Exercise 6.6

Consider a world in which there are no sentient beings, and nothing else that could introduce norms or laws. Since there are no norms at this world, one might hold that nothing is obligatory relative to the world's norms, and nothing is permitted. Explain why this casts doubt on the validity of (Dual1) and (Dual2) in the logic of relativist obligation and permission.

### Exercise 6.7

A system of norms is *intolerant* if it requires of itself that it is in force and does not allow any other norms. That is, if the norms at  $w$  are intolerant, then only worlds with the same norms conform to these norms. Show that the relativist logic of intolerant norms validates (4) and (5).

## 6.3 Norms and circumstances

We have assumed that something ought to be the case iff it is the case at all worlds where no (relevant) norms are violated. On closer inspection, many ordinary statements about oughts and obligations do not fit this analysis.

Suppose you are walking past a drowning baby. You ought to rescue the baby. But are you rescuing 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 example shows that even on an absolutist approach, obligations and permissions can vary from world to world. In worlds where you are passing by a drowning baby, you

are obliged to save it. In other worlds, you are not. The relevant (moral) norms may well be the same in either case. What varies are your circumstances.

In general, what is required or permitted usually depends not just on the norms, but also on the circumstances – for example, on what you are able to do, and on what consequences the available options would have.

We can account for this dependence on the circumstances by changing our interpretation of the accessibility relation. Previously, we assumed that a world  $v$  is accessible from  $w$  iff all the norms (or all the norms at  $w$ ) are respected at  $v$ . Let's add another condition: relevant circumstances at  $w$  must also obtain at  $v$ . For example, if  $w$  is a world at which you come across a drowning baby then any accessible world must also be a world at which you come across a drowning baby. In all ideal worlds *among these*, you rescue the baby.

Here is the redefined accessibility relation, in terms of which we might try to analyse O and P:

A world  $v$  is deontically accessible from a world  $w$  iff (a)  $v$  is circumstantially accessible from  $w$ , and (b) no norms (at  $w$ ) are violated at  $v$ .

The parenthetical '(at  $w$ )' must be included on a relativist approach, but not on an absolutist approach.

We might want to say more about the circumstantial accessibility relation in clause (a). Recall that a world  $v$  is circumstantially accessible from  $w$  if relevant circumstances that obtain in  $w$  also obtain in  $v$ . Often, the "relevant circumstances" that we seem to hold fixed when we reason about norms comprise everything that is *settled*, in the sense of section 1.5 – everything 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; the falling is a "relevant circumstance" that takes place at every world accessible from  $w$ . Arguably, however, there are cases in which we treat worlds as accessible that aren't open. 'Jones ought to be here', for example, can be true even if it's settled that Jones is somewhere else. Perhaps the circumstantial accessibility relation that figures in clause (a) varies with conversational context.

With the new definition of deontic accessibility,  $OA$  says that among the circumstantially accessible worlds, all ideal worlds are  $A$ -worlds. We could make this more explicit. Let ' $N$ ' be a propositional constant whose intended meaning is that all norms are satisfied. We can then use  $\Box(N \rightarrow A)$  to express that  $A$  is required, where the box expresses the relevant kind of circumstantial necessity. This approach to formalizing obligation statements goes back to Leibniz.

**Exercise 6.8**

How could we define  $P$  in terms of  $\Box$  and  $N$ , so that  $P$  is the dual of  $O$ ?

**Exercise 6.9**

Show that the Leibnizian approach renders the (U)-schema valid, assuming that the circumstantial accessibility is reflexive. You have to first translate the schema into the Leibnizian language.

Whichever language we use to express it, our revised concept of obligation has a serious problem. It assumes that the circumstantially accessible worlds include ideal worlds, at which no norms are violated. For suppose there are no such worlds. Then no world is deontically accessible! We would have to say that everything is required and nothing permitted (because all instances of  $OA$  are true and all instances of  $PA$  false at worlds that can't access anything).

Now remember that we don't assume that all worlds are ideal. If a world is not ideal, then it is hard to see why the worlds that are circumstantially accessible from it should always include ideal worlds. Couldn't the "relevant circumstances" that are held fixed include some norm violations?

The problem is brought about by Arthur Prior's *Samaritan Paradox*. Suppose someone (Smith) has been injured in a robbery, and Jones has the opportunity to help. We want to say that Jones ought to help the victim: he helps the victim at all deontically accessible worlds. But then the robbery must have taken place at all these worlds. (In a world without a robbery, there is no victim to help.) Here, the circumstantially accessible worlds all contain a violation of norms. In a truly ideal world, nobody would have been robbed and nobody would be in need of help.

We need to adjust our revised definition of deontic accessibility. How could we do that?

In the Samaritan Paradox, the robbery is settled; it has happened at all worlds that are compatible with the "relevant circumstances". None of these worlds are ideal. Crucially, however, worlds at which Jones doesn't help the victim are even *worse*, in terms of norm violations, than worlds at which he helps the victim. Both kinds of worlds are non-ideal, because the victim got robbed. But our norms don't just divide the possible worlds into ideal and non-ideal; they allow for finer distinctions among non-ideal worlds. Jones ought to help the victim because that's what he does in the *best* worlds among those he can bring about.

These considerations suggest that we should redefine deontic accessibility as follows, to properly account for the dependence of obligations and permissions on circumstances.

A world  $v$  is deontically accessible from a world  $w$  iff  $v$  is among the best worlds (by the norms at  $w$ ) among those that are circumstantially accessible from  $w$ .

As before, the parenthetical ‘(by the norms at  $w$ )’ would be needed in a relativist account and not in an absolutist account.

It can be useful to factor out the circumstantial and deontic components that enter into the new definition. I don’t mean to separate them in the formal language, as in Leibniz’s proposal. Rather, I mean to separate them in the definition of a model.

Let’s define a new type of model. Instead of a deontic accessibility relation, we have two ingredients besides the worlds  $W$  and the interpretation function  $V$ . One is a circumstantial accessibility relation. The other is a world-relative “order” that tells us which worlds are better than others, relative to the norms at any given world (which may be the norms at every world, on an absolutist approach).

Let ‘ $u <_w v$ ’ mean that world  $u$  is better than world  $v$  relative to the norms at  $w$ . The symbol ‘ $<$ ’ hints at the idea that  $u$  contains *fewer* violations of norms than  $v$ . We assume that for any world  $w$ , the relation  $<_w$  is transitive. We also assume that it is asymmetric, meaning that if  $u <_w v$  then it is not the case that  $v <_w u$ . Asymmetric and transitive relations are known as **strict partial orders**.

### Definition 6.1

A **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 strict partial order  $<_w$  on  $W$  (the world-relative ranking of worlds as better or worse), and
- a function  $V$  that assigns to each sentence letter of  $\mathcal{L}_M$  a subset of  $W$ .

Now we need to say under what conditions a sentence of the form  $OA$  is true at a world in an ordering model. Informally,  $OA$  will 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  $S$  and any partial order  $<$ , let  $\text{Min}^<(S)$  be the set of  $<$ -minimal

members of  $S$ :

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

An expression of the form ' $\{x : \dots x \dots\}$ ' denotes the set of all things  $x$  that satisfy the condition  $\dots x \dots$ . So  $\text{Min}^<(S)$  is the set of all things  $v$  that are members of  $S$  and for which there are no members  $u$  of  $S$  for which  $u < v$ .

Here, then, are the truth-conditions for  $OA$  and  $PA$  in deontic ordering models:

**Definition 6.2: Ordering semantics**

If  $M$  is a deontic ordering model and  $w$  a world in  $M$ , then

$M, w \models OA$  iff  $M, v \models A$  for all  $v \in \text{Min}^{<_w}(\{u : wRu\})$

$M, w \models PA$  iff  $M, v \models A$  for some  $v \in \text{Min}^{<_w}(\{u : wRu\})$

This is just a formal way of saying that  $OA$  is true at  $w$  iff  $A$  is true at the best worlds (by the norms at  $w$ ) among the worlds that are circumstantially accessible at  $w$ .

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

The logic of obligation and permission now depends on formal properties of the circumstantial accessibility relation  $R$  and the deontic orderings  $<_w$ . In section 1.5, I argued that the logic of historical necessity (of what is settled and open) is S5. This suggests that in normal contexts,  $R$  is an equivalence relation. If we adopt an absolutist approach, on which the orderings  $<_w$  are the same for every world  $w$ , we then still get KD45. If we allow the orderings to vary from world to world, we still get D, unless we impose further restrictions on the orderings.

**Exercise 6.10**

Amy ought to promise to help Betty or to help Carla. She doesn't make either promise. If she had promised to help Betty, she would be obligated to help Betty. If she had promised 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, since Amy makes neither promise, she is neither obligated to help Betty nor to help Carla. Explain why this casts doubt on the assumption that deontic accessibility is euclidean.

**Exercise 6.11**

Suppose fatalism is true and the only world that is open (circumstantially accessible) relative to any world  $w$  is  $w$  itself. Can you describe the resulting deontic logic (on either an absolutist or a relativist approach)?

Ordering models prove useful when we want to formalize statements with modal operators and if-clauses, like (1)–(3).

- (1) If you smoke then you must smoke outside.
- (2) If you miss the deadline for tax returns then you must pay a fine.
- (3) If you have promised to call your parents then you must call them.

How would you translate these into our language  $\mathcal{L}_M$ ? You seem to face a choice between (W) and (N).

(W)  $O(p \rightarrow q)$

(N)  $p \rightarrow Oq$

In (W), the operator  $O$  is said to have **wide scope** because it applies to the entire conditional  $p \rightarrow q$ . In (N), the operator has **narrow scope** because it only applies to the consequent  $q$ .

On reflection, neither translation is satisfactory. Starting with (N), note that  $p \rightarrow Oq$  and  $\neg Oq$  together entail  $\neg p$ . But from (1), together with the assumption that you are not required to smoke ( $\neg Oq$ ), we surely can't infer that you do not in fact smoke.

(W) is not much better. For one, in our Kripke-style semantics,  $O(p \rightarrow q)$  is entailed by  $O(\neg p)$ . But it is easy to imagine a scenario in which you must not smoke, or you must submit your tax return before the deadline, but in which (1) and (2) are false.

Both (N) and (W) would also license a problematic form of “strengthening the antecedent”. For example, they both suggest that (3) entails (4).

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

**Exercise 6.12**

Give tree proofs with the K-rules to show that  $p \rightarrow Or$  entails  $(p \wedge q) \rightarrow Or$ , and that  $O(p \rightarrow r)$  entails  $O((p \wedge q) \rightarrow r)$ .

Let's think about what is expressed by statements like (1)–(4). Intuitively, when we ask what must be done if  $p$  is the case, we are limiting our attention to situations in which  $p$  is the case, and consider which of *these* situations best conform to the relevant norms. It is irrelevant whether  $p$  is in fact the case or whether it ought to be the case. (1) says – roughly – that among worlds where you smoke, the best worlds are worlds where you smoke outside. Worlds where you smoke inside are worse than worlds where you smoke outside. Similarly for (2). A world at which you miss the deadline for tax returns and pay the fine contains only one violation of the tax rules. Worlds at which you miss the deadline and don't pay the fine contain two. The best worlds among those at which you miss the deadline are worlds at which you pay the fine. Likewise for (3). Among worlds at which you have promised to call your parents, the best are worlds at which you keep the promise and call them.

The if-clause in sentences like (1)–(3) therefore seems to restrict the worlds over which the modal operator quantifies. Whereas 'ought  $q$ ' alone says that  $q$  is true at the best of the open worlds, 'if  $p$  then ought  $q$ ' says that  $q$  is true at the best of the open worlds *at which  $p$  is true*.

There is no way to express these truth-conditions with the resources of  $\mathcal{L}_M$ . But we can introduce a new, binary operator for **conditional obligation**. The operator is often written ' $O(\cdot/\cdot)$ ', with a slash separating the two argument places. Intuitively,  $O(B/A)$  means that  $B$  ought to be the case if  $A$  is the case.

The formal truth-conditions for  $O(B/A)$  are much like those for  $OB$ , except that we add the assumption  $A$  to the circumstances that are held fixed:

**Definition 6.3: Ordering semantics for conditional obligation**

If  $M$  is a deontic ordering model and  $w$  a world in  $M$ , then

$M, w \models O(B/A)$  iff  $M, v \models B$  for all  $v \in \text{Min}^{<_w}(\{u : wRu \text{ and } M, u \models A\})$ .

Here,  $\{u : wRu \text{ and } M, u \models A\}$  is the set of worlds  $u$  that are circumstantially accessible from  $w$  and at which  $A$  is true.  $\text{Min}^{<_w}(\{u : wRu \text{ and } M, u \models A\})$  is the set that comprises the best of these worlds. So  $O(B/A)$  is true at  $w$  iff  $B$  is true at all of the best  $A$ -worlds that are accessible at  $w$ .

### Exercise 6.13

“Deontic detachment” is the inference from  $OA$  and  $O(B/A)$  to  $OB$ . “Factual detachment” is the inference from  $A$  and  $O(B/A)$  to  $OB$ . Which of these are valid on the present semantics?

### Exercise 6.14

In exercise 6.1, you were asked to translate the following statements.

- (c) Jones ought to help his neighbours.
- (d) If Jones is going to help his neighbours, then he ought to tell them he’s coming.
- (e) If Jones isn’t going to help his neighbours, then he ought to not tell them he’s coming.

Let’s add a fourth statement:

- (f) Jones is not going to help his neighbours.

Intuitively, none of these four statements is entailed by one of the others. Moreover, they don’t impose contradictory requirements on Jones. This shows that your translations in exercise 6.1 were incorrect. Explain. (This puzzle is due to Roderick Chisholm.)

### Exercise 6.15

The dual of conditional obligation is conditional permission. Spell out truth-conditions for  $P(B/A)$  that parallel the truth-conditions I have given for  $O(B/A)$ , so that  $P(B/A)$  is equivalent to  $\neg O(\neg B/A)$ .

## 6.4 Further challenges

Many apparent problems for standard deontic logic arise from the dependence of obligations on circumstances. We can avoid these problems by using deontic ordering models and formalizing conditional obligation statements with the binary  $O(\cdot/\cdot)$  operator. There are, however, other problems and “paradoxes” for which this move doesn’t help. I will mention three.

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. 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 logic K,  $O p$  and  $O \neg p$  entail  $O A$ , 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.

Another family of problems arises from the fact that in any logic defined in terms of Kripke models,  $O$  is closed under logical consequence, meaning that if  $O A$  is true and  $A$  entails  $B$ , then  $O B$  is true. Since logical truths are logically entailed by everything, it follows that all logical truths come out as obligatory. (This is easy to see semantically. A logical truth is true at all worlds; so it is true at all deontically accessible worlds.) But ought it to be the case that it either rains or doesn't rain?

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,  $\neg O(p \vee \neg p)$  entails  $P\neg(p \vee \neg p)$ . 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 permissible 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 the victim 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 all the deontic logics we have so far considered, the claim that Jones ought to stop the victim's blood flow therefore entails that the victim ought to be bleeding. But wouldn't it be better if the victim weren't 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 the victim is bleeding. We are interested in what should be done *given* the state in which Jones found the victim. Worlds where the victim isn't injured are set aside; they are not circumstantially accessible. But circumstantial accessibility can shift with conversational context. The claim that the victim ought to be bleeding is pointless if we hold fixed the victim's state of injury. So when we evaluate *this* claim, we naturally assume that the relevant circumstantial accessibility relation does not hold fixed the injuries. Intuitively, we are no longer considering what

should be done given the state in which Jones found the victim, but whether that state itself should have obtained. 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,  $O(A \vee B)$  does not entail  $PA \wedge PB$ . (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,  $P(A \vee B)$  does not entail  $PA \wedge PB$ .

We could add the missing principles.

$$(R) \quad O(A \vee B) \rightarrow (PA \wedge PB)$$

$$(FC) \quad P(A \vee B) \rightarrow (PA \wedge PB)$$

But both of these have unacceptable consequences when added to the minimal modal logic K. With the help of (R), we could show that  $OA$  entails  $PB$ :  $OA$  entails  $O(A \vee B)$ , which by (R) entails  $PA \wedge PB$ . But clearly ‘you ought to mail the letter’ does not entail ‘you may burn the letter’. Similarly for (FC). In K,  $PA$  entails  $P(A \vee B)$ ; by (FC),  $P(A \vee B)$  entails  $PB$ . But ‘you may have beer’ does not entail ‘you may have wine’.

#### Exercise 6.16

Analogous puzzles to those raised by Ross’s Paradox and the Paradox of Free Choice arise for epistemic ‘must’ and ‘might’. Can you give examples?

## 6.5 Neighbourhood semantics

In reaction to apparent problems for standard deontic logic, some have argued that we should not interpret obligation and permission in terms of quantification over possible worlds. If we give up this core tenet of Kripke semantics, we can define “non-normal”

logics weaker than K. (A **normal** modal logic is a modal logic that can be defined in terms of classes of Kripke frames.)

A 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. Validity is still defined as truth at all worlds in all (suitable) 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. Thus  $N(w)$  is a set of sets of worlds. Each set of world in  $N(w)$  is necessary at  $w$ .

#### Definition 6.4

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  $\mathcal{L}_M$  a subset of  $W$ .

The interpretation of non-modal sentences at neighbourhood models works just as 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 proxy for the proposition expressed by  $A$ . Then:

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

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. The neighbourhood function  $N$  can easily make  $p$  necessary at a world without making  $p \vee q$  necessary, even though  $p$  entails  $p \vee q$ . 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 the victim’s injuries even though it is not the case that someone 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 no longer follows that any proposition whatsoever is obligatory.

We may further hope to escape the problems from section 6.3 that led us to introduce a primitive conditional obligation operator. I argued that the wide-scope translation  $O(A \rightarrow B)$  of conditional obligation sentences is problematic because  $O(A \rightarrow B)$  is entailed by  $O(\neg A)$ . In neighbourhood semantics, this entailment fails.

Bare neighbourhood semantics determines a very weak logic called **E**. It is axiomatized by (Dual), (CPL), and a rule (called “RN”) that allows inferring  $\Box A \leftrightarrow \Box B$  from  $A \leftrightarrow B$ . We can get stronger logics, with more validities, by imposing conditions on the neighbourhood function  $N$ .

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 permitted. By the neighbourhood semantics for **P**,  $A$  is permitted 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; in that case,  $\neg A$  is true at no worlds, and  $[\neg A]^M$  is the empty set. If we want logical truths to be permitted, we therefore have to stipulate that  $N(w)$  never contains the empty set.

In Kripke semantics, the assumption that logically necessary truths are permitted is equivalent to the assumption that (every instance of) the (D)-schema  $O A \rightarrow P A$  is valid. Both assumptions correspond to seriality of the accessibility relation. In neighbourhood semantics, we can distinguish between the two assumptions. While the permissibility of logical truths requires that  $N(w)$  doesn’t contain the empty set, the validity of  $O A \rightarrow P A$  requires that  $N(w)$  doesn’t contain 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  $(\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$  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 6.17

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

For some purposes, even the minimal logic of neighbourhood semantics is too strong.

Return to the intuitive “Free Choice” principle from the previous section:

$$(FC) \quad P(A \vee B) \rightarrow (P A \wedge P B)$$

We have seen that this principle is untenable in Kripke semantics. 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  $P A$  and  $P B$  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  $[A]^M$  is in  $N(w)$  iff  $[B]^M$  is in  $N(w)$ . Likewise,  $[\neg A]^M$  is in  $N(w)$  iff  $[\neg B]^M$  is in  $N(w)$ .

Now any sentence  $A$  is logically equivalent to  $(A \wedge B) \vee (A \wedge \neg B)$ , for any  $B$ . In the logic E,  $P A$  therefore entails  $P((A \wedge B) \vee (A \wedge \neg B))$ . By (FC),  $P((A \wedge B) \vee (A \wedge \neg B))$  entails  $P(A \wedge B)$ . We could still reason from ‘you may have a cookie’ to ‘you may have a cookie and burn down the house’.

#### Exercise 6.18

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 one can’t give a Kripke semantics for this concept of belief. (Although one can give a neighbourhood semantics.) *Hint:* One rule of probability says that if  $p$  and  $q$  are independent propositions, then the probability of their conjunction  $p \wedge q$  is the product of their individual probabilities.



## 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 we read  $r$  as 'it is raining', we will use  $F r$  to express that it will be raining at some point in the future. We use  $P r$  to express that it has been raining at some point in the past. In general:

$F A$  is true at a time  $t$  iff  $A$  is true at some time after  $t$ .

$P A$  is true at a time  $t$  iff  $A$  is true at some time before  $t$ .

The operators  $F$  and  $P$  can be nested. We can use  $F P r$  to express that at some point it will have rained,  $P F r$  to say that it was once going to rain,  $P P r$  to say that there was a time before which it rained, and  $F F r$  to say that there will come a time after which it will rain.

Unlike  $\Box$  and  $\Diamond$ ,  $F$  and  $P$  are not duals of each other:  $\neg P A$  is not equivalent to  $F \neg A$ , and  $\neg F A$  is not equivalent to  $P \neg A$ . But it is useful to have duals of  $F$  and  $P$ . We therefore introduce two more operators.  $G$  will be the dual of  $F$ , and  $H$  the dual of  $P$ .

Intuitively,  $G A$  means that  $A$  is always *going* to be the case. (Hence the symbol 'G'.) If it is not the case that at some point in the future it will not rain ( $\neg F \neg r$ ), then it is always going to be the case that it will rain ( $G r$ ). Similarly,  $H A$  means that  $A$  *has* always been the case. If it is not the case that at some point in the past it was not raining ( $\neg P \neg r$ ), then it has always been raining ( $H r$ ).

We can state the truth-conditions of  $G A$  and  $H A$  in parallel to the above truth-conditions for  $F A$  and  $P A$ :

$G A$  is true at a time  $t$  iff  $A$  is true at all times after  $t$ .

$H A$  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  $F, P, G, H$  is known as the **language of basic temporal logic**. We will sometimes call it  $\mathcal{L}_t$ .

### Exercise 7.1

Translate the following sentences into the language of basic temporal logic.

- (a) It has never been warm.
- (b) There will be a sea battle.
- (c) There will not have been a sea battle.
- (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

A complete scenario for temporal logic needs to tell us what times there are, how they are ordered, and what is going on at each of them. We can represent such a scenario, together with an interpretation of  $\mathcal{L}_t$ 's 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. This is enough to determine, for every  $\mathcal{L}_t$ -sentence and every time, whether the sentence is true at that time.

### 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  $\mathcal{L}_T$  a subset of  $T$ .

We use ‘ $M, t \models A$ ’ as a short-hand notation to express that sentence  $A$  is true at time  $t$  in model  $M$ . The following definition formally specifies the truth-value of every  $\mathcal{L}_T$ -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  $T$ ,  $P$  is any sentence letter, and  $A, B$  are any  $\mathcal{L}_T$ -sentences, then

- |     |                                    |   |
|-----|------------------------------------|---|
| (a) | $M, t \models P$                   | iff $t$ is in $V(P)$ .  |
| (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 $M, t \not\models A$ or $M, t \models B$ .                              |
| (f) | $M, t \models A \leftrightarrow B$ | iff $M, t \models (A \rightarrow B)$ and $M, t \models (B \rightarrow A)$ . |
| (g) | $M, t \models FA$                  | iff $M, s \models A$ for some $s \in T$ such that $t < s$ .                 |
| (h) | $M, t \models GA$                  | iff $M, s \models A$ for all $s \in T$ such that $t < s$ .                  |
| (i) | $M, t \models PA$                  | iff $M, s \models A$ for some $s \in T$ such that $s < t$ .                 |
| (j) | $M, t \models HA$                  | 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 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.

All this should remind you of our Kripke semantics for  $\mathcal{L}_M$  in chapter 3. In fact, temporal models *are* Kripke models, as defined on page 49. I have merely relabelled the set ' $W$ ' as ' $T$ ', and the relation ' $R$ ' as ' $<$ '. Definition 7.2 resembles definition 3.2 from page 49, 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  $H$ ). 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 use the set of natural numbers 0,1,2, etc. Let's say that the precedence relation  $<$  holds between  $t$  and  $s$  iff  $t$  is less than  $s$ . So  $0 < 1$  and  $1 < 25$ . (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 'less than'.) Finally, let's say that the interpretation function assigns to  $p$  the set of all even numbers.

Let's call this model  $M$ . By definition 7.2, we can figure out the following facts, among others.

- $M, 0 \models p$  (because 0 is even);

- $M, 0 \models F p$  (because there are even numbers greater than 0);
- $M, 0 \models G F p$  (because for every number there is a greater number that is even);
- $M, 0 \models \neg F G p$  (because there is no number for which all greater numbers are even).

### Exercise 7.2

Now let  $M$  be the following model. As before,  $T$  is the set of natural numbers  $\{0, 1, 2, \dots\}$ , and  $t < s$  iff  $t$  is less than  $s$ . This time,  $V(p)$  is the set of numbers less than 10. Which of the following statements are true?

- (a)  $M, 0 \models F p \wedge F \neg p$
- (b)  $M, 0 \models G \neg p$
- (c)  $M, 0 \models F G \neg p$
- (d)  $M, 0 \models G F p$
- (e)  $M, 0 \models G(F p \rightarrow F F p)$
- (f)  $M, 0 \models F H p$
- (g)  $M, 0 \models \neg P(p \vee \neg p)$
- (h)  $M, 0 \models H p$

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 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  $s$  iff the computation can lead from  $t$  to  $s$ . If the computation 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 other applications, 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-later 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 might have the right formal structure.

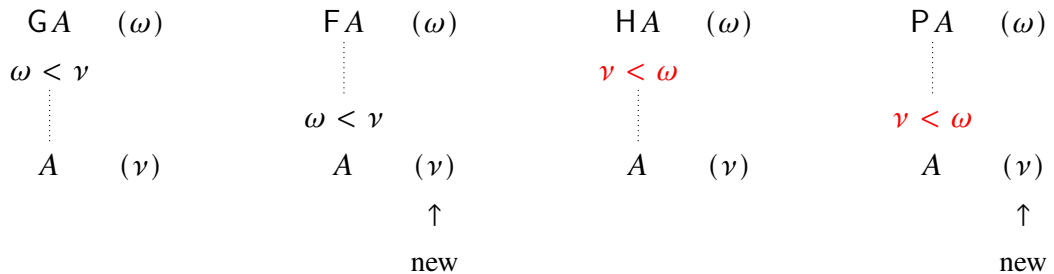
For yet other applications, we may want to assume that time is **dense**, meaning that whenever  $t < s$  then there is another point of time lying between  $t$  and  $s$ . This assumption is common in physics. The natural numbers, by contrast, have a **discrete** structure. There is no natural number between 2 and 3. For dense models, we could use real or rational numbers (fractions) instead of natural numbers.

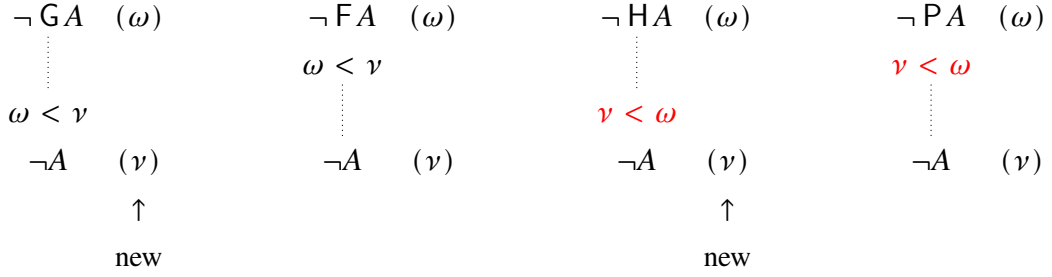
If we want to take seriously what physics tells us about time, it is not enough to assume that time is dense. We also need to reconceptualize the set  $T$ . According to the theory of special relativity, whether a point in time is earlier or later than another is relative to a spatial frame of reference. 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;  $\langle x_1, x_2, x_3, t \rangle < \langle y_1, y_2, y_3, s \rangle$  holds iff the second point can be reached from the first without travelling faster than the speed of light.

### 7.3 Logics of time

Let's define the minimal temporal logic  $K_t$  as the set of  $\mathcal{L}_t$ -sentences that are true at all times in all temporal models. Since temporal models are just Kripke models, proof methods for the minimal modal logic  $K$  are easily adapted to  $K_t$ . The main novelty is that the rules for the box and the diamond can be used twice over, once for the forward-looking operators  $F$  and  $G$ , and once for the backward-looking  $P$  and  $H$ .

In the tree method for  $K_t$ , we have all the  $K$ -rules, with  $G$  as the box and  $F$  as the diamond. In addition, we have rules for  $H$  as the box and  $P$  as the diamond with a reversed perspective on the accessibility (or precedence) relation:





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  $H$ :

$$(GK) \quad G(A \rightarrow B) \rightarrow (G A \rightarrow G B)$$

$$(HK) \quad H(A \rightarrow B) \rightarrow (H A \rightarrow H B)$$

We also have two versions of Necessitation, and two versions of (Dual):

$$(GDI) \quad \neg F A \leftrightarrow G \neg A$$

$$(HDI) \quad \neg P A \leftrightarrow H \neg A$$

$$(GNec) \quad \text{If } A \text{ occurs in a proof, } G A \text{ may be appended.}$$

$$(HNec) \quad \text{If } A \text{ occurs in a proof, } H A \text{ may be appended.}$$

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  $H$ :

$$(Con1) \quad A \rightarrow G P A$$

$$(Con2) \quad A \rightarrow H F A$$

These axioms and rules, added to those of classical propositional logic, define an axiomatic calculus that is sound and complete for  $K_t$ . (Completeness is easily proved with the canonical model technique.)

### Exercise 7.3

Show with the help of definition 7.2 that all instances of (Con1) and (Con2) are true at all times in all temporal models.

### Exercise 7.4

Give  $K_t$ -tree proofs for the following schemas.

- (a)  $A \rightarrow G P A$
- (b)  $A \rightarrow H F A$
- (c)  $F A \rightarrow H F F A$
- (d)  $P G A \rightarrow P F A$
- (e)  $H A \leftrightarrow H F H A$

For most applications,  $K_t$  is too weak. We will want to impose further restrictions on the relevant temporal models. 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 assume that the precedence relation is transitive. This corresponds to the (4)-schema for  $G$ . It also corresponds to the (4)-schema for  $H$ .

$$(4G) \quad G A \rightarrow G G A$$

$$(4H) \quad H A \rightarrow H H A$$

### Exercise 7.5

Explain why, if a relation  $<$  is transitive, then so is its converse. The converse  $>$  of  $<$  is the relation that holds between  $x$  and  $y$  iff  $y < x$ .

Another plausible condition is that no time is earlier than itself. Formally,  $<$  should be *irreflexive*, so that no element of  $T$  is  $<$ -related to itself. We know that reflexivity corresponds to the (T)-schema, whose (forward-looking) temporal analogue would be  $G A \rightarrow A$ . 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. The left-to-right direction is implied by the answer to exercise 4.8. But we can give a more direct argument.

Suppose that some sentence  $A$  is not valid in the class of all frames. We show that  $A$  is not valid in the class of irreflexive frames. That  $A$  is not valid in the class of all

frames means that there is some world  $w$  in some model  $M = \langle W, R, V \rangle$  at which  $A$  is false. We will 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  $w$ : if  $w \in V(P)$  then  $w' \in V(P)$ ;
- $w'$  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  $M^i$ .  $\square$

Given transitivity, irreflexivity is closely related to asymmetry. Recall from the previous chapter that  $<$  is asymmetric if whenever  $t < s$  then not  $s < t$ . There is no modal schema that corresponds to asymmetry.

#### Exercise 7.6

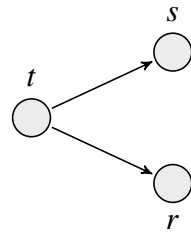
Show that a transitive relation is irreflexive iff it is asymmetric.

#### Exercise 7.7

A popular idea in many cultures is that time is circular. Does this cast doubt on asymmetry? What about irreflexivity?

In the previous chapter, I mentioned that transitive and irreflexive relations are called (strict) partial orders. The name reflects the fact that such orders need not order everything. In a model of branching time, for example, we can have  $t < s$  and  $t < r$  but neither  $s < r$  nor  $r < s$ ; in that case,  $r$  and  $s$  are not ordered by the precedence relation.

We can rule out such cases by imposing the requirement of **connectedness**, also known as *completeness* or *totality*. This demands



that for any points  $t$  and  $s$ , either  $t < s$  or  $t = s$  or  $s < t$ . An irreflexive, transitive, and connected relation is called a **(strict) linear order** (or a *strict total order*).

For some applications, we may want linearity in only one direction. Many philosophers have been attracted to a branching-future conception of time, where a point in time may have more than one future, but only one past. In such models, we would only require **backward-linearity**: that if  $s < t$  and  $r < t$ , then either  $s < r$  or  $s = r$  or  $r < s$ .

An axiom schema corresponding to backward-linearity is (BL):

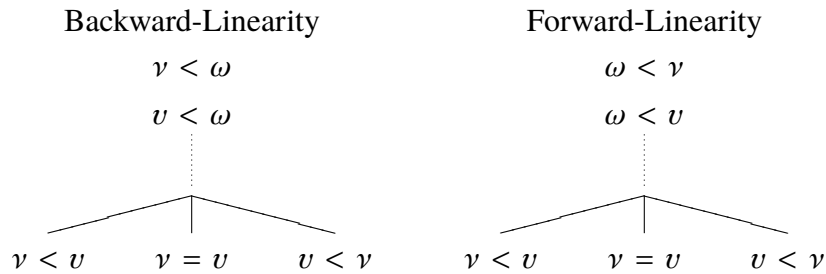
$$(BL) \quad FPA \rightarrow (FA \vee A \vee PA)$$

Forward-linearity – the assumption that if  $t < s$  and  $t < r$ , then either  $s < r$  or  $s = r$  or  $r < s$  – corresponds to (FL):

$$(FL) \quad PFA \rightarrow (PA \vee A \vee FA)$$

The conjunction of (BL) and (FL) is valid on a frame iff the frame's precedence relation does not branch in either direction. This is not quite the same as connectedness, because it allows for frames with parallel time lines. There is no schema that corresponds to connectedness.

The tree rules for backward-linearity and forward-linearity directly reflect the definition of the two properties.



These rules create *three* branches. The “identity nodes”  $\nu = u$  on the middle branch state that two world/time labels refer to the same thing. (This must be taken into account when reading off countermodels from open branches.) We need two further rules to deal with identity nodes. Both of these rules are called ‘Identity’.



### Exercise 7.8

Use the tree method to check which of the following sentences are valid, assuming time is linear (i.e., using the Transitivity, Backward-Linearity, Forward-Linearity, and Identity rules).

- (a)  $(Fp \wedge Fq) \rightarrow F(p \wedge q)$
- (b)  $P G G p \rightarrow G G p$
- (c)  $P F p \rightarrow (P p \vee (p \vee F p))$
- (d)  $P H p \rightarrow H p$
- (e)  $F G p \rightarrow G F p$
- (f)  $F(Gq \wedge \neg p) \rightarrow G(p \rightarrow (Gp \rightarrow q))$
- (g)  $(P(q \wedge Hq) \wedge P G q) \rightarrow H q$

The precedence relation in relativistic spacetime is neither backward-linear nor forward-linear. But it has a weaker property: 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 a common later 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:

- (FG)  $F G A \rightarrow G F A$
- (PG)  $P H A \rightarrow H P A$

### Exercise 7.9

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 7.10

Show that the schema  $FA \rightarrow FFA$  corresponds to density. (You have to show that (a) whenever a frame is dense then  $FA \rightarrow FFA$  is valid on the frame, and (b) whenever  $FA \rightarrow FFA$  is valid on a frame then the frame is dense.)

### Exercise 7.11

Can you find an  $\mathcal{E}_T$ -expression stating that  $p$  is true at all times? Can you do so if you make assumptions about the precedence relation?

## 7.4 Branching time

In section 1.5 we looked at the idea that the future is “open” while the past is “settled”. We can still influence (say) whether we will exercise tomorrow, but not whether we have exercised yesterday. Some have argued that the openness of the future calls for a non-linear model of time, with multiple branches into the future. On one branch, we would exercise tomorrow, on another we would not.

This line of thought appears to conflate temporal and modal considerations. The precedence relation in models of time is normally understood as a purely temporal relation – as the earlier-later relation. The fact that we can bring about a world in which we exercise tomorrow and a world in which we don’t exercise does not entail that both kinds of tomorrow take place here in the actual world.

If we want to make explicit the connections between settledness and time, it is better to use a multi-modal language with circumstantial operators for settledness and openness in addition to the purely temporal operators  $F, G, P, H$ . We could then say things like  $PA \rightarrow \Box PA$  to formalize the claim that if  $A$  has happened then it is settled that  $A$  has happened.

There are nonetheless good reasons to consider branching models of time. I already mentioned that such models are widely used in computer science, where the “times” represent states of a computational process and the precedence relation has a semi-modal interpretation, holding between two states iff the first can lead to the second. I also mentioned that the precedence relation in relativistic spacetime allows for branching, although diverging spacetime branches ultimately reconverge. A more classical form of branching (without reconvergence) has been argued to follow from a certain 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.

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 concern other times are reducible to more fundamental truths about the present. If, for example, 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 that entail (and explain) yesterday's sea battle. Perhaps the relevant facts about the present specify the distribution of physical particles and fields etc. together with the general laws of nature. If the laws are deterministic, then the complete truth about the present distribution of particles and fields etc. together with the laws fixes all truths about the past and about the future. But suppose the laws 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 regard both of these futures as equally actual.

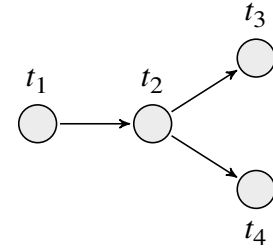
Let's assume, then, that we want to reason about branching time. This is less straightforward than it might at first appear.

Two pieces of terminology will be useful. First, let's define a **history** in a model  $\langle T, <, V \rangle$  as a maximal linearly ordered subset of  $T$ . That is, a history is a collection of times  $H$  such that

- (i) for all  $t$  and  $s$  in  $H$ , either  $t < s$  or  $t = s$  or  $s < t$ , and
- (ii) no further member of  $T$  could be added to  $H$  without making (i) false.

The model (or rather, frame) depicted on the right contains two histories:  $\{t_1, t_2, t_3\}$  and  $\{t_1, t_2, t_4\}$ .

For the second piece of terminology, let  $t$  be any time in any model. Any maximal linearly ordered set of times *later than*  $t$  will be called a **future of**  $t$ . In the model 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,  $Gp$  is true at  $t$  iff  $p$  is true at all times *in all futures of*  $t$ ;  $Fp$ , 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 is often thought to be problematic if we want  $Fp$  to translate 'it will be the case that  $p$ '.

To illustrate, suppose I'm about to toss a coin. In one future (let's assume), the coin will land heads, in another it will land tails. By definition 7.2, both  $Fh$  and  $Ft$  are true. But should we say that the coin will land heads and also that it will land tails?

We could adopt an alternative semantics for  $F$  according to which  $Fp$  is true at  $t$  iff  $p$  is true at some time in *all* futures of  $t$ :

$$M, t \models FA \text{ iff every future of } t \text{ contains some } s \text{ such that } M, s \models A.$$

This is known as the **Peircean interpretation** of  $F$  (after Charles S. Peirce; the name is due to Arthur Prior).

On the Peircean account,  $Fp$  is false whenever  $p$  only takes place in one of several futures. If we keep the classical interpretation of  $G$ , both  $Fp$  and  $G\neg p$  can be false; the two operators 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 7.12

Explain why the Peircean interpretation renders  $p \rightarrow HFp$ , an instance of (Con2), invalid.

A different line 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 history. Histories are linear by definition, so the problems raised by multiple futures disappear. To say that  $p$  is the case in *some* history, or in *all* histories, Ockhamists add new operators  $\Diamond$  and  $\Box$  that quantify over histories. The Peircean  $F$  operator is equivalent to  $\Box F$  in Ockhamism.  $\Box Fp$  says that every future contains a time at which  $p$  is true;  $\Diamond Fp$ , by contrast, would say that some future contains a time where  $p$  is true.

Here is the full Ockhamist semantics.

### Definition 7.3: Ockhamist Semantics

If  $M = \langle T, <, V \rangle$  is a temporal model,  $H$  is a history in  $M$ ,  $t$  is a member of  $H$ ,  $P$  is any sentence letter, and  $A, B$  are any sentences in the Ockhamist language, then

- |     |                                       |   |
|-----|---------------------------------------|---|
| (a) | $M, H, t \models P$                   | iff $t$ is in $V(P)$ .  |
| (b) | $M, H, t \models \neg A$              | iff $M, H, t \not\models A$ .   |
| (c) | $M, H, t \models A \wedge B$          | iff $M, H, t \models A$ and $M, H, t \models B$ .                                 |
| (d) | $M, H, t \models A \vee B$            | iff $M, H, t \models A$ or $M, H, t \models B$ .                                  |
| (e) | $M, H, t \models A \rightarrow B$     | iff $M, H, t \not\models A$ or $M, H, t \models B$ .                              |
| (f) | $M, H, t \models A \leftrightarrow B$ | iff $M, H, t \models (A \rightarrow B)$ and $M, H, t \models (B \rightarrow A)$ . |
| (g) | $M, H, t \models F A$                 | iff $M, H, s \models A$ for some $s$ in $H$ such that $t < s$ .                   |
| (h) | $M, H, t \models G A$                 | iff $M, H, s \models A$ for all $s$ in $H$ such that $t < s$ .                    |
| (i) | $M, H, t \models P A$                 | iff $M, H, s \models A$ for some $s$ in $H$ such that $s < t$ .                   |
| (j) | $M, H, t \models H A$                 | iff $M, H, s \models A$ for all $s$ in $H$ such that $s < t$ .                    |
| (k) | $M, H, t \models \Box A$              | iff $M, J, t \models A$ for all histories $J$ that contain $t$ .                  |
| (l) | $M, H, t \models \Diamond A$          | iff $M, J, t \models A$ for some history $J$ that contains $t$ .                  |

A sentence is *valid* in Ockhamist semantics if it is true at all times  $t$  on all histories  $H$  (containing  $t$ ) in all models. As always, we can get stronger conceptions of validity – stronger logics – by adding further constraints on the precedence relation.

### Exercise 7.13

Which of the following schemas are valid in Ockhamist semantics?

- (a)  $\Box A \rightarrow A$
- (b)  $\Box A \rightarrow \Box \Box A$
- (c)  $\Diamond A \rightarrow \Box \Diamond A$
- (d)  $\Box F A \rightarrow F \Box A$
- (e)  $P A \rightarrow \Box P \Diamond A$

There is something odd about the Ockhamist approach. 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  $F p$  true in this scenario (under the given interpretation of  $p$ )? What about  $F(p \vee \neg p)$ ? Or  $Gp \rightarrow GGp$ ?

Ockhamism refuses to give an answer. In Ockhamism, sentences are only true or false relative to a model and a time *and a history*. A branching-time scenario, however, does not fix a particular history. 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 history that contains a sea battle,  $F p$  is true. Relative to other histories,  $F p$  is false.

If we insist that logical validity should formalize the idea of truth in all scenarios under all interpretations of non-logical vocabulary then we can't accept the official definition of validity in Ockhamist semantics. We have to extend the Ockhamist semantics to specify under what conditions a sentence is true *in a model at a time*, without fixing a history. Then we can say that a sentence is valid iff it is true at all times in all models.

A simple way to do this is to stipulate that a sentence is true at time in an (Ockhamist) model iff it is true relative to *all* histories that contain the time:

$$M, t \models A \text{ iff } M, H, t \models A \text{ for all histories } H \text{ that contain } t.$$

This is known as a **supervaluationist** semantics.

Supervaluationism is often used when a formal semantics defines truth relative to an “extra” parameter that doesn't correspond to any feature of a conceivable scenario. In Ockhamist semantics, that parameter is the history  $H$ . For a different application, consider vagueness. If  $p$  translates ‘it is warm’, and the temperature is borderline warm, it is not clear what we should say about the truth-value of  $p$ , and about various complex sentences containing  $p$ . One popular approach to vagueness is to first define truth relative to a *sharpening* of vague expressions. Relative to a sharpening on which temperatures above 15.0 degrees Celsius are warm,  $p$  has a clear truth-value in any conceivable scenario, as do complex sentences containing  $p$ . Since an actual scenario does not fix a particular sharpening, this semantics contains an extra parameter. We can define a notion of truth without that parameter by saying that a sentence is true in a scenario iff it is true in that scenario relative to every eligible sharpening.

Supervaluationist accounts tend to have some non-classical features. Suppose we live in a branching world in which one future contains a sea battle and another doesn't. Let  $p$  express that a sea battle takes place. According to supervaluationist Ockhamism, neither  $Fp$  nor  $\neg Fp$  is true in that scenario. Both are true relative to some but not relative to all histories. So neither is simply true. Assuming that a sentence is *false* if its negation is true,  $Fp$  is neither true nor false!

Logics in which a sentence can have a third status besides (mere) truth and (mere) falsity are called **three-valued**. Three-valued approaches to branching time are sometimes defended by the intuition that if a sea battle occurs on some but not all branches of the future, then one can't truly assert that a battle *will* occur nor that it *won't* occur.

The Polish logician Jan Łukasiewicz argued that statements about the future are either true, false, or “indeterminate”. To accommodate this third truth-value, he proposed three-valued truth-tables specifying how the truth-value of complex sentences are determined by the truth-value of their parts. For example, he suggested that if two sentences  $A$  and  $B$

are indeterminate, then their conjunction  $A \wedge B$ , disjunction  $A \vee B$ , and negations  $\neg A$ ,  $\neg B$  are also indeterminate.

In the sea battle scenario, Łukasiewicz’s account renders  $F s \vee \neg F s$  indeterminate, assuming  $F s$  is indeterminate. This is often regarded as problematic: even if we shouldn’t assert that there will be a sea battle, it is argued that we are justified to assert that there either will or there won’t be a sea battle. The supervaluationist form of Ockhamism, while also three-valued, avoids this problem. On the supervaluationist interpretation,  $F s$  and  $\neg F s$  are neither true nor false in the sea battle scenario, but  $F s \vee \neg F s$  is true.

#### Exercise 7.14

Let’s say that a sentence is *super-valid* if it is true at all times in all models, where truth at a time in a model is understood in accordance with supervaluationist Ockhamism. Explain why the super-valid sentences are precisely the sentences that are valid by the original Ockhamist definition of validity (just below definition 7.3).

#### Exercise 7.15

Things are more complicated for entailment. Let’s say that  $A$  *Ockham-entails*  $B$  iff there is no time on any history in any temporal model at which  $A$  is true and  $B$  false. Let’s say that  $A$  *super-entails*  $B$  iff there is no time in any temporal model at which  $A$  is true and  $B$  false, where truth at a time in a model is defined in accordance with supervaluationism. Is Ockham-entailment equivalent to super-entailment? Explain.

## 7.5 Extending the language

The expressive resources of standard modal and temporal 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 history 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 ‘ $\circ$ ’). Informally,  $\mathbf{X}A$  means that  $A$  is true at the next point in time. Formally:

$M, t \models \mathbf{X}A$  iff  $M, s \models A$  for some  $s$  such that (i)  $t < s$  and (ii)  $s < r$  for all  $r$  such that  $r \neq s$  and  $t < r$ .

With the help of  $\mathbf{X}$ , we can also say that  $A$  is true in two units of time ( $\mathbf{X}\mathbf{X}A$ ), in three units of time ( $\mathbf{X}\mathbf{X}\mathbf{X}A$ ), and so on. The corresponding operator for talking about the *previous* point in time is usually written  $\mathbf{Y}$ .

A more powerful extension of  $\mathcal{L}_T$  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,  $A \mathbf{S} B$  is true iff  $B$  was true at some time in the past and  $A$  has always been true since then;  $A \mathbf{U} B$  is true iff  $B$  will be true at some time in the future and  $A$  will always be true until then. Formally:

$M, t \models A \mathbf{S} B$  iff there is some  $s$  with  $s < t$  for which  $M, s \models B$ , and for all  $r$  with  $s < r < t$ , we have  $M, r \models A$ .

$M, t \models A \mathbf{U} B$  iff there is some  $s$  with  $t < s$  for which  $M, s \models B$ , and for all  $r$  with  $t < r < s$ , we have  $M, r \models A$ .

The operators  $\mathbf{F}$ ,  $\mathbf{G}$ ,  $\mathbf{P}$ , and  $\mathbf{H}$  can all be defined in terms of  $\mathbf{S}$  and  $\mathbf{U}$ . For example,  $\mathbf{P}A$  is equivalent to  $(p \vee \neg p) \mathbf{S} A$ . And  $\mathbf{F}A$  is equivalent to  $(p \vee \neg p) \mathbf{U} A$ .

#### Exercise 7.16

Define  $\mathbf{X}A$  in terms of  $\mathbf{U}$ .

Another noteworthy addition to temporal logic is the “Now” operator  $\mathbf{N}$ . To see the point of this operator, consider the following multi-modal statement.

- (3) We already knew yesterday that there would be a test today.

Using  $\mathbf{Y}$  for ‘yesterday’, we might try to translate (3) as  $\mathbf{Y}Kp$ , where  $p$  translates ‘there is a test’. But that’s wrong. By the semantics for  $\mathbf{Y}$ ,  $\mathbf{Y}Kp$  is true today iff  $Kp$  is true yesterday (using days as temporal units). Since  $Kp$  entails  $p$ , it follows that  $\mathbf{Y}Kp$  is true today only if  $p$  is true *yesterday*. But the test takes place today, not yesterday.

Intuitively, the problem is that ‘today’ in (3) refers to the present day, even though it occurs in the scope of the ‘yesterday’ operator. 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  $N$ , we can translate (3) as  $\forall K N p$ , and (4) as  $F \forall x (N R x \rightarrow P x)$ . (We will have a closer look at quantified modal logic in later chapters.)

Intuitively, the  $N$  operator allows us to look outside the scope of an embedding operator.  $P N p$ , for example, is true if there is some time in the past such that  $p$  is true not at that time, but at the present. How does this work formally?

By the semantics of  $P$ ,

$$M, t \models P N p \text{ iff } M, s \models N p \text{ for some time } s < t.$$

Now we want  $M, s \models N p$  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 simplest way to achieve this is to define truth relative to pairs of times. One of the times is shifted by the temporal 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  $T$ ,  $P$  is any sentence letter, and  $A, B$  are any  $\mathcal{L}_T$ -sentences, then

- (a)  $M, t_0, t \models P$  iff  $t$  is in  $V(P)$ .
- (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  $M, t_0, t \not\models A$  or  $M, t_0, t \models B$ .
- (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)  $M, t_0, t \models F A$  iff  $M, t_0, s \models A$  for some  $s$  in  $T$  such that  $t < s$ .
- (h)  $M, t_0, t \models G A$  iff  $M, t_0, s \models A$  for all  $s$  in  $T$  such that  $t < s$ .
- (i)  $M, t_0, t \models P A$  iff  $M, t_0, s \models A$  for some  $s$  in  $T$  such that  $s < t$ .
- (j)  $M, t_0, t \models H A$  iff  $M, t_0, s \models A$  for all  $s$  in  $T$  such that  $s < t$ .
- (k)  $M, t_0, t \models N A$  iff  $M, t_0, t_0 \models A$ .

Like the Ockhamist semantics from the previous section, this semantics has an extra parameter. An ordinary scenario is represented by a single time in a model, not by a pair of times. So we need to specify under what conditions a sentence is true at a (single) time. Here, the standard approach is not supervaluation but “diagonalization”:

$$M, t \models A \text{ iff } M, t, t \models A.$$

This “two-dimensional” semantics correctly predicts that  $P N p$  entails  $p$ .

1. Assume  $M, t \models P N p$ .
2. Then  $M, t, t \models P N p$ , by the definition of truth at a time in a model.
3. Then  $M, t, s \models N p$  for some  $s < t$ , by clause (i) of definition 7.4.
4. Then  $M, t, t \models p$ , by clause (k) of definition 7.4.
5. Then  $M, t \models p$ , 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. For example,  $N p \rightarrow p$  is valid, in the sense that it is true at all times in all models. But  $G(N p \rightarrow p)$  is invalid. If  $p$  is true at  $t$  and false at some time after  $t$ , then  $G(N p \rightarrow p)$  is false at  $t$ . So we must give up the forward and backward Necessitation rules. The fact that something is logically true does not entail that it will always be true!

#### Exercise 7.17

‘It might have been that everyone who is actually rich is poor.’ This says that there is a world  $w$  such that everyone who is rich *at the actual world* is poor *at  $w$* . To formalize statements like these, we need a modal operator analogous to  $N$  that takes us back to the actual world, even in the scope of other modal operators. This operator is called the *actually* operator. Let’s write it as  $A$  and add it to  $\mathcal{L}_M$ . Can you find a sentence  $B$  in this language that is logically true but not necessarily true, in the sense that  $B$  is true at all worlds in all models but  $\Box B$  is not?



## 8 Conditionals

### 8.1 Material conditionals

We are often interested not just in whether something is in fact the case, but also in whether it is (or would be) the case *if* something else is (or would be) the case. We might, for example, wonder in what will happen to the climate if we don't reduce greenhouse gases, or whether World War 2 could have been avoided if certain steps had been taken in the 1930s.

A sentence stating that something is (or would be) the case if something else is (or would be) the case is called a **conditional**. What exactly, do these statements mean? What is their logic? Philosophers have puzzled over these questions for more than 2000 years, with no agreement in sight.

One attractively simple view is that a conditional 'if  $A$  then  $B$ ' is true iff the antecedent  $A$  is false or the consequent  $B$  is true. This would make 'if  $A$  then  $B$ ' equivalent to 'not  $A$  or  $B$ '. Conditionals with these truth-conditions are called **material conditionals**.

The conditionals  $A \rightarrow B$  of classical logic are material.  $A \rightarrow B$  is equivalent to  $\neg A \vee B$ . The attractively simple view that English conditionals are material conditionals would mean that we can faithfully translate English conditionals into  $\mathcal{L}_M$ -sentences of the form  $A \rightarrow B$ . Is this correct?

There are some arguments for a positive answer. Suppose I make the following promise.

- (1) If I don't have to work tomorrow then I will help you move.

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, you could not fault me for breaking 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 true iff  $A$  is false or  $B$  is true.

Another argument for analysing English conditionals as material conditionals 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* inference.) Suppose I tell you that Nadia is either in Rome or in Paris. Trusting me, you can infer that if she's not in Rome then she's in Paris. Now we can reason as follows.

To begin, suppose that  $A$  and ‘if  $A$  then  $B$ ’ are both true. Plausibly, we can infer that  $B$  is true as well: *modus ponens* is valid for English conditionals. This means that if  $A$  is true and  $B$  is false, then ‘if  $A$  then  $B$ ’ is false. Now suppose, alternatively, that  $A$  is false or  $B$  is true. Then ‘not- $A$  or  $B$ ’ is true. By or-to-if, we can infer that ‘if  $A$  then  $B$ ’ is true as well. Thus ‘if  $A$  then  $B$ ’ is true iff  $A$  is false or  $B$  is true.

Despite these arguments, most philosophers and linguists don’t think that English conditionals are material conditionals. Consider these facts about logical consequence (in classical propositional logic).

- (M1)  $B \models A \rightarrow B$
- (M2)  $\neg A \models A \rightarrow B$
- (M3)  $\neg(A \rightarrow B) \models A$
- (M4)  $A \rightarrow B \models \neg B \rightarrow \neg A$
- (M5)  $A \rightarrow B \models (A \wedge C) \rightarrow B$

If English conditionals were material conditionals then the following inferences, corresponding to (M1)–(M5), would be valid.

- (E1) There won’t be a nuclear war. Therefore: If Russia attacks the US with nuclear weapons then there won’t be a nuclear war.
- (E2) There won’t be a nuclear war. Therefore: If there will be a nuclear war then nobody will die.
- (E3) It is not the case that if it will rain tomorrow then the Moon will fall onto the Earth. Therefore: It will rain tomorrow.
- (E4) 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.
- (E5) 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.

These inferences do not sound good. If we wanted to defend the view that English conditionals are material conditionals we would have to explain why they sound bad even though they are valid. We will not explore this option any further.

**Exercise 8.1**

Can you find a different analysis of English conditionals that, like the material analysis, would make conditionals truth-functional, but that would render all of (E1)–(E5) invalid?

Even those who defend the material analysis of English conditionals admit that it does not work for all English conditionals. Consider (2).

- (2) If water is heated to 100° C, it evaporates.

This shouldn't be translated as  $p \rightarrow q$ . Intuitively, (2) states that *in all (normal) cases* where water is heated to 100° C, it evaporates. It is a quantified, or modal claim.

Another important class of conditionals that can't be analysed as material conditionals 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*; 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**. Intuitively, an indicative conditional states that something is *in fact* the case on the assumption that something else is the case. A subjunctive conditional like (4) states that something *would be* the case if something else *were* the case. Typically we know that the "something else" is not in fact the case. We know, for example, that Shakespeare wrote *Hamlet* and therefore that the antecedent of (4) is false. For this reason, subjunctive conditionals are also called *counterfactual conditionals* or simply *counterfactuals*.

It should be clear that subjunctive conditionals are not material conditionals. I said that (4) is almost certainly false. But it almost certainly has a false antecedent. So the corresponding material conditional is almost certainly true.

## 8.2 Strict conditionals

One apparent difference between material conditionals  $A \rightarrow B$  and conditionals in natural language is that  $A \rightarrow B$  requires no connection between the antecedent  $A$  and the consequent  $B$ . Consider (1).

- (1) If we leave after 5, we will miss the train.

Intuitively, someone who utters (1) wants 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 that are equivalent to  $\Box(A \rightarrow B)$  are called **strict conditionals**. The label goes back to C.I. Lewis (1918), who also introduced the abbreviation  $A \rightarrow B$  for  $\Box(A \rightarrow B)$ .

Lewis was not interested in ‘if ...then ...’ sentences. He introduced  $A \rightarrow B$  to formalize ‘ $A$  implies  $B$ ’ or ‘ $A$  entails  $B$ ’. His intended use of  $\rightarrow$  roughly matches our use of the double-barred turnstile ‘ $\models$ ’. But there are important differences. The turnstile is an operator in our *meta-language*; Lewis’s  $\rightarrow$  is an *object-language* operator that, like  $\wedge$  or  $\rightarrow$ , can be placed between any two sentences in a formal language to generate another sentence in the language.  $p \rightarrow (q \rightarrow p)$  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  $p \rightarrow q$  is true on some interpretation of the sentence letters and false on others. If  $p$  means that it is raining heavily and  $q$  that it is raining, then  $p \rightarrow q$  is true because the hypothesis that it is raining heavily implies that it is raining.

Let’s set aside Lewis’s project of formalizing the concept of implication. Our goal is to find an object-language construction that functions like ‘if ...then ...’ in English. To see whether ‘...  $\rightarrow$  ...’ can do the job, let’s have a closer look at the logic of strict conditionals.

Since  $A \rightarrow B$  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  $A$  is false at  $v$  or  $B$  is true at  $v$ . We therefore have the following truth-conditions for strict conditionals.

**Definition 8.1: Kripke semantics for  $\rightarrow$**

If  $M = \langle W, R, V \rangle$  is a Kripke model, then

$M, w \models A \rightarrow B$  iff for all  $v$  such that  $wRv$ , either  $M, v \not\models A$  or  $M, v \models B$ .

**Exercise 8.2**

$A \rightarrow B$  is equivalent to  $\Box(A \rightarrow B)$ . Can you fill the blank in: ‘ $\Box A$  is equivalent to —’, using no modal operator other than  $\rightarrow$ ?

As always, the logic of strict conditionals depends on what constraints we put on the accessibility relation. Without any constraints,  $\rightarrow$  does not validate *modus ponens*, in the sense that  $A \rightarrow B$  and  $A$  together do not entail  $B$ . We can see this by translating  $A \rightarrow B$  back into  $\Box(A \rightarrow B)$  and setting up a tree. Recall that to test whether some premises entail a conclusion, we start the tree with the premises and the negated conclusion.

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, node 1 can’t be expanded, so there is nothing more we can do.

**Exercise 8.3**

Give a countermodel in which  $p \rightarrow q$  and  $p$  are true at some world while  $q$  is false.

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  $\rightarrow$  is valid on all and only the reflexive frames. Reflexivity is precisely what we need to render *modus ponens* valid. Since *modus ponens* looks plausible for English conditionals (as I’ve argued on p. 146), we’ll probably want the relevant Kripke models to be reflexive.

**Exercise 8.4**

Using the tree method, and translating  $A \rightarrow B$  into  $\Box(A \rightarrow B)$ , confirm that following claims hold, for all  $A, B, C$ .

- (a)  $\models_K A \rightarrow A$
- (b)  $A \rightarrow B \models_K \neg B \rightarrow \neg A$
- (c)  $A \rightarrow B \models_K (A \wedge C) \rightarrow B$
- (d)  $A \rightarrow B, B \rightarrow C \models_K A \rightarrow C$
- (e)  $(A \vee B) \rightarrow C \models_K (A \rightarrow C) \wedge (B \rightarrow C)$
- (f)  $A \rightarrow (B \rightarrow C) \models_T (A \wedge B) \rightarrow C$
- (g)  $A \rightarrow B \models_{S4} C \rightarrow (A \rightarrow B)$
- (h)  $((A \rightarrow B) \rightarrow C) \rightarrow (A \rightarrow B) \models_{S5} A \rightarrow B$

Which of these schemas do you think should be valid if we assume that  $A \rightarrow B$  translates indicative conditionals ‘if  $A$  then  $B$ ’?

We could now look at other conditions on the accessibility relation and decide whether they should be imposed, based on what they would imply for the logic of conditionals. But let’s take a shortcut.

I have suggested that sentence (1) might be understood as saying 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. There are, for example, worlds at which the train departs two hours later, worlds at which we live right next to the station, and so on. When I say that it is impossible to leave after 5 and still make it to the train, I arguably mean that it is impossible *given what we know about the departure time, our location, etc.*

Generalizing, a tempting proposal is that the accessibility relation that is relevant for indicative conditionals like (1) is the epistemic accessibility relation that we studied in chapter 5, where a world  $v$  is accessible from  $w$  iff it is compatible with what is known at  $w$ . On that hypothesis, the logic of indicative conditionals is determined by the logic of epistemic necessity. We don’t need to figure out the relevant accessibility relation 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.

Sly Pete and Mr. Stone are playing poker on a Mississippi 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. He also knows that Pete would not call unless he has the better hand. So among the worlds compatible with Zack's knowledge, all worlds at which Pete calls are worlds at which Pete wins. If  $p$  translates 'Pete called' and  $q$  'Pete won', then  $p \rightarrow q$  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. Among the worlds compatible with Jack's knowledge, all worlds at which Pete calls are therefore worlds at which Pete loses. Relative to Jack's information state,  $p \rightarrow \neg q$  is true.

Another advantage of the "epistemically strict" interpretation is that it might explain why indicative conditionals with antecedents that are known to be false seem defective. For example, imagine a scenario in which Jones has gone to work. In that scenario, is (2) true or false?

- (2) If Jones has not gone to work then 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 Jones has gone to work, it is unclear how we are meant to assess whether Jones is helping his neighbours *if* he has not gone to work. On the epistemically strict interpretation, (2) says that Jones is helping his neighbours at all epistemically accessible worlds at which Jones hasn't gone to work. Since we know that Jones has gone to work, there are no epistemically accessible worlds at which he hasn't gone to work. And if there are no  $A$ -worlds then we naturally balk at the question whether all  $A$ -worlds are  $B$ -worlds. (In logic, we resolve to treat 'all  $A$ s are  $B$ ' as true if there are no  $A$ s. Accordingly, (2) comes out true on the epistemically strict analysis. But we can still explain why it seems defective.)

We have found a promising alternative to the hypothesis that indicative conditionals are material conditionals. According to the present alternative, they are epistemically strict conditionals – 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. That’s why the conditional is true. At all accessible worlds at which Shakespeare didn’t write *Hamlet*, someone else wrote the play. When we evaluate the subjunctive conditional – ‘If Shakespeare hadn’t written *Hamlet*, then someone else would have’ – we do consider worlds at which *Hamlet* was never written, even though we know that the actual world is not of that kind. If subjunctive conditionals are strict conditionals, then their accessibility relation 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 could track.

This is one problem for the strict analysis of natural-language conditionals. Another problem lies in the logic of strict conditionals. Remember (E1)–(E5) from page 146. If English conditionals are strict conditionals, then (E1)–(E3) are invalid. For example, while  $q$  entails  $p \rightarrow q$ , it does not entail  $p \rightarrow q$ . But the strict analogs of (M4) and (M5) still hold, no matter what we say about accessibility (see exercise 8.4):

$$A \rightarrow B \models \neg B \rightarrow \neg A;$$

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

So we still predict that the inferences (E4) and (E5) are valid.

- (E4) 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.
- (E5) 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.

#### Exercise 8.5

The badness of (E4) and (E5) suggests that indicative conditionals can’t be analysed as strict conditionals. Can you give a similar argument suggesting that *subjunctive* conditionals can’t be analysed as strict conditionals?

#### Exercise 8.6

A plausible norm of pragmatics is that a sentence should only be asserted if it is known to be true. Let’s call a sentence *assertable* if it is known to be true. Show

that if the logic of knowledge is at least S4, then an epistemically strict conditional  $A \rightarrow B$  is assertable iff the corresponding material conditional  $A \rightarrow B$  is assertable.

### Exercise 8.7

Explain why the ‘or-to-if’ inference from ‘ $p$  or  $q$ ’ to ‘if not  $p$  then  $q$ ’ is invalid on the assumption that the conditional is epistemically strict. How could a friend of this assumption explain why the inference nonetheless looks reasonable, at least in normal situations? (Hint: Remember the previous exercise.)

## 8.3 Variably strict conditionals

Let’s have a closer look at subjunctive conditionals. As I am writing these notes, I am sitting in Coombs Building, room 2228, with my desk facing the wall to Al Hájek’s office in room 2229. In light of these facts, (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.

There is no logical connection between the antecedent of (1) and the consequent. There are many possible worlds at which I drill a hole through the wall behind my desk and don’t reach 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 (1) is a strict conditional then all such worlds must be inaccessible.

Now consider (2).

- (2) If the office spaces had been randomly reassigned yesterday then Al’s office would (still) be next to mine.

(2) seems false, or at least very unlikely. But if (2) is a strict conditional, and worlds at which Al is not in room 2229 or I am not in 2228 are inaccessible – as they seem to be for (1) – then (2) should be true. Among 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 at which Al’s office is next to mine. When we evaluate (2), it looks like we no longer hold fixed who is in which office. Worlds that were inaccessible for (1) are accessible for (2).

So the accessibility relation, at least for subjunctive conditionals, appears to vary from conditional to conditional. As David Lewis put it, subjunctive conditionals seem to be 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 in the 1960s.)

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 these worlds deviate in some minimal way to allow the event to take place. Afterwards the worlds unfold in accordance with the general laws of the actual world.

For example, if we wonder what would have happened if Shakespeare hadn't written *Hamlet*, we are interested in worlds that are like the actual world until 1599, at which point some mundane circumstances prevent Shakespeare from writing *Hamlet*. We are not interested in worlds at which Shakespeare was never born, or in which the laws of nature are radically different from the laws at our world. One might reasonably judge that Shakespeare would have been a famous author even if he hadn't written *Hamlet*, although we would hardly be famous in worlds in which he was never born.

Likewise for (1). Here we are considering 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. Worlds where I'm not in room 2228 can be 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.

This suggests that 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. The closest worlds (to the actual world) at which Shakespeare didn't write *Hamlet* are worlds that almost perfectly match the actual world until 1599, then deviate a little so that Shakespeare didn't write *Hamlet*, and afterwards still resemble the actual world with respect to the general laws of nature. We will not try to spell out in full generality what the relevant closeness measure should look like.

Let ' $v <_w u$ ' mean 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 the world-relative ordering  $<$ .

1. If  $v <_w u$  then  $u \not<_w v$ . (Asymmetry)
2. If  $v <_w u$ , then for all  $t$  either  $v <_w t$  or  $t <_w u$ . (Quasi-connectedness)
3. For any non-empty set of worlds  $X$  and world  $w$  there is a  $v$  in  $X$  such that there is no  $u$  in  $X$  with  $u <_w v$ .

Asymmetric and quasi-connected relations are known as **weak orders**. Asymmetry is self-explanatory. Quasi-connectedness is more often called *negative transitivity*, because

it is equivalent to the assumption that if  $t \not\prec s$  and  $s \not\prec r$  then  $t \not\prec r$ . It ensures that the “equidistance” relation that holds between  $v$  and  $u$  if neither  $v \prec_w u$  nor  $u \prec_w v$  is an equivalence relation. With these two assumptions, we can picture each world  $w$  as associated with nested spheres of worlds;  $v \prec_w u$  means that  $v$  is in a more narrow  $w$ -sphere than  $u$ .

Assumption 3 is known as the **Limit Assumption**. It 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 8.8

Show that asymmetry and quasi-connectedness imply transitivity.

#### Exercise 8.9

Define  $\leq_w$  so that  $v \leq_w u$  iff  $u \not\prec_w v$  (that is, iff it is not the case that  $u \prec_w v$ ). Informally,  $v \leq_w u$  means that  $v$  is at least as similar to  $w$  in the relevant respects as  $u$ . Many authors use  $\leq$  rather than  $<$  as their basic notion. Can you express the above three conditions on  $<$  in terms of  $\leq$ ?

We are going to introduce a variably strict operator  $\Box \rightarrow$  so that  $A \Box \rightarrow B$  is true at a world  $w$  iff  $B$  is true at the closest worlds to  $w$  at which  $A$  is true. Models for a language with the  $\Box \rightarrow$  operator must contain closeness orderings  $<$  on the set of worlds.

#### Definition 8.2

A **similarity model** consists of

- a non-empty set  $W$ ,
- for each  $w$  in  $W$  a weak order  $<_w$  that satisfies the Limit Assumption, and
- a function  $V$  that assigns to each sentence letter a subset of  $W$ .

To formally state the semantics of  $\Box \rightarrow$ , we can re-use a concept from section 6.3. Let  $S$  be an arbitrary set of worlds, and let  $w$  be some world (that may or may not be in  $S$ ). It will be useful to have an expression that picks out the most similar worlds to  $w$ , among all the worlds in  $S$ . This expression is  $\text{Min}^{<_w}(S)$ , which we have defined as follows in

section 6.3:

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

Now  $\{u : M, u \models A\}$  is the set of worlds (in model  $M$ ) at which  $A$  is true. So  $\text{Min}^{<_w}(\{u : M, u \models A\})$  is the set of those  $A$ -worlds that are closest to  $w$ . We want  $A \Box \rightarrow B$  to be true at  $w$  iff  $B$  is true at the closest  $A$ -worlds to  $w$ .

**Definition 8.3: Similarity semantics for  $\Box \rightarrow$**

If  $M$  is a similarity model and  $w$  a world in  $M$ , then

$M, w \models A \Box \rightarrow B$  iff  $M, v \models B$  for all  $v$  in  $\text{Min}^{<_w}(\{u : M, u \models A\})$ .

You may notice that  $A \Box \rightarrow B$  works almost exactly like  $O(B/A)$  from section 6.3. There, I said that for any world  $w$  in any deontic ordering model  $M$ ,

$$M, w \models O(B/A) \text{ iff } M, v \models B \text{ for all } v \text{ in } \text{Min}^{<_w}(\{u : wRu \text{ and } M, u \models A\}).$$

The main difference is that conditional obligation is sensitive to an accessibility relation. If that relation is an equivalence relation then this makes no difference to the logic.

Of course, the order  $<$  in deontic ordering models is supposed to represent degree of conformity to norms, while the order  $<$  in similarity models represents a certain similarity ranking in the evaluation of subjunctive conditionals. A different type of ordering might be in play when we evaluate indicative conditionals, which some have argued should also be interpreted as variably strict. But again, these differences in interpretation don't affect the logic.

Suppose we add the  $\Box \rightarrow$  operator to the language of standard propositional logic. The set of sentences in this language that are true at all worlds in all similarity models is known as **system V**. There are tree rules and axiomatic calculi for this system, but they aren't very user-friendly. We will only explore the system semantically.

To begin, we can check whether *modus ponens* is valid for  $\Box \rightarrow$ . That is, we check whether the truth of  $A$  and  $A \Box \rightarrow B$  at a world in a similarity model entails the truth of  $B$ .

Assume that  $A$  and  $A \Box \rightarrow B$  are true at a world  $w$ . By definition 8.3, the latter means that  $B$  is true at all the closest  $A$ -worlds to  $w$  (at all worlds in  $\text{Min}^{<_w}(\{u : M, u \models A\})$ ). The world  $w$  itself is an  $A$ -world. If we could show that  $w$  is among the closest  $A$ -worlds to itself then we could infer that  $A$  is true at  $w$ .

Without further assumptions, however, we can't show this. If we want to validate *modus ponens*, we must add a further constraint on our models: that every world is

among the closest worlds to itself. More precisely,

for all worlds  $w$  and  $v$ ,  $v \not\prec_w w$ .

This assumption is known as **Weak Centring**. The logic we get if we impose this constraint is **system VC**.

### Exercise 8.10

Should we accept Weak Centring for deontic ordering models?

### Exercise 8.11

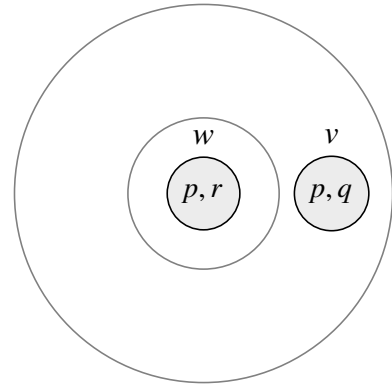
Explain why  $A \Box \rightarrow B$  entails  $A \rightarrow B$ , assuming Weak Centring.

### Exercise 8.12

Show that if  $A$  is true at no worlds, then  $A \Box \rightarrow B$  is true.

None of the problematic inferences (E1)–(E5) are valid if the relevant conditionals are interpreted as variably strict. (E5), for example, would assume that  $p \Box \rightarrow r$  entails  $(p \wedge q) \Box \rightarrow r$ . But it does not. We can give a countermodel with two worlds  $w$  and  $v$ ;  $p$  is true at both worlds,  $q$  is true only at  $v$ , and  $r$  only at  $w$ ; if  $w$  is closer to itself than  $v$ , then  $p \Box \rightarrow r$  is true at  $w$  (because the closest  $p$ -worlds to  $w$  are all  $r$ -worlds), but  $(p \wedge q) \Box \rightarrow r$  is false at  $w$  (because the closest  $(p \wedge q)$ -worlds to  $w$  aren't all  $r$ -worlds).

The diagram on the right represents this model. The circles around  $w$  depict the similarity spheres.  $w$  is closer to  $w$  than  $v$  because it is in the innermost sphere around  $w$ , while  $v$  is only in the second sphere. (If  $v$  were also in the innermost sphere then the two worlds would be equally close to  $w$ . That's allowed.) In general, we can represent the assumption that a world  $v$  is closer to a world  $w$  than a world  $u$  ( $v \prec_w u$ ) by putting  $v$  in a closer sphere around  $w$  than  $u$ . I have not drawn any spheres around  $v$  because it doesn't matter what these look like.



### Exercise 8.13

Draw countermodels showing that (E1)–(E4) are invalid if the conditionals are translated as statements of the form  $A \Box \rightarrow B$ . (Hint: You never need more than two worlds.)

The logic of variably strict conditionals is weaker than the logic of strict conditionals. Some have argued that it is too weak to explain our reasoning with conditionals. It is, for example, not hard to see that the following statements are all false. (The corresponding statements for  $\neg 3$  are true; see exercise 8.4.)

1.  $p \Box \rightarrow q, q \Box \rightarrow r \models p \Box \rightarrow r$
2.  $((p \vee q) \Box \rightarrow r) \models (p \Box \rightarrow r) \wedge (q \Box \rightarrow r)$
3.  $p \Box \rightarrow (q \Box \rightarrow r) \models (p \wedge q) \Box \rightarrow r$

If English conditionals are variably strict, this means (for example) that we can't infer 'if  $p$  then  $r$ ' from 'if  $p$  then  $q$ ' and 'if  $q$  then  $r$ '. But isn't this a valid inference?

Well, perhaps not. Stalnaker gave the following counterexample, using cold-war era subjunctive conditionals.

If J. Edgar Hoover had been born a Russian, he would be a communist.

If Hoover were a communist, he would be a traitor.

Therefore, if Hoover had been born a Russian, he would be a traitor.

### Exercise 8.14

Can you find a case where 'if  $p$  or  $q$  then  $r$ ' does not appear to entail 'if  $p$  then  $r$ ' and 'if  $q$  then  $r$ '? You can use either indicative or subjunctive conditionals. (Hint: Try to find a case in which 'if  $p$  or  $q$  then  $p$ ' sounds acceptable.)

The semantics I have presented for  $\Box \rightarrow$  is a middle ground between that of Lewis and Stalnaker. Stalnaker assumes that  $<_w$  is not just quasi-connected, but connected: for any  $w, v, u$ , either  $v <_w u$  or  $v = u$  or  $u <_w v$ . (' $v = u$ ' means that  $v$  and  $u$  are the same world.) This rules out ties in similarity: no sphere contains more than one world.

Stalnaker's logic (called **C2**) is stronger than Lewis's VC. The following principle of "Conditional Excluded Middle" is C2-valid but not VC-valid:

$$(CEM) \quad (A \Box \rightarrow B) \vee (A \Box \rightarrow \neg B)$$

Whether conditionals in natural language satisfy Conditional Excluded Middle is a matter of ongoing debate. On the one hand, it is natural to think that ‘it is not the case that if  $p$  then  $q$ ’ entails ‘if  $p$  then not  $q$ ’, which suggests that the principle is valid. 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 of the coins? Arguably, I might have gotten heads and I might have gotten tails. Either result is possible, but neither *would* have come about.

### Exercise 8.15

Explain why the following statements are true, for all  $A, B, C$ :

- (a)  $A \wedge B \models_{C2} A \Box \rightarrow B$
- (b)  $A \Box \rightarrow (B \vee C) \models_{C2} (A \Box \rightarrow B) \vee (A \Box \rightarrow C)$

Lewis not only rejects connectedness, but also the Limit Assumption. He argued that there might 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  $A \Box \rightarrow B$  is true. That does not seem right. Lewis therefore gives a more complicated semantics:

$M, w \models A \Box \rightarrow B$  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  $u <_w v$ ,  $M, w \models A \rightarrow B$ .

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 Restrictors

Consider these two statements.

- (1) If it rains we always stay inside.
- (2) If it rains we sometimes stay inside.

On its most natural reading, (1) says that we stay inside at all times at which it rains. We can express this in  $\mathcal{E}_M$ , using the box as a universal quantifier over the relevant times. (So  $\Box A$  now means ‘always  $A$ ’.) The translation would be  $\Box(r \rightarrow s)$ .

One might expect that (2) should then be translated as  $\Diamond(r \rightarrow s)$ , where the diamond is an existential quantifier over the relevant times (‘sometimes’). But  $\Diamond(r \rightarrow s)$  is equivalent

to  $\Diamond(\neg r \vee s)$ . This is true whenever  $\Diamond\neg r$  is true. (2), however, isn't true simply because it doesn't always rain. On its most salient reading, (2) says there are times at which it rains *and* we stay inside. Its correct translation is  $\Diamond(r \wedge s)$ .

This is a little surprising, given that (2) seems to contain a conditional. Does the conditional here express a conjunction?

Things get worse if we look at (3).

(3) If it rains we usually stay inside.

Let's introduce an operator  $M$  for 'usually', so that  $MA$  is true at a time iff  $A$  is true at *most* times. Can you translate (3) with the help of  $M$ ?

You can't. Neither  $M(r \rightarrow s)$  nor  $M(r \wedge s)$  capture the intended meaning of (3).  $M(r \wedge s)$  entails that  $r$  is usually true. But (3) doesn't entail that it usually rains.  $M(r \rightarrow s)$  is true as long as  $r$  is usually false, even if we're always outside when it is raining. You could try to bring in some of the new kinds of conditional that we've encountered in the previous sections. How about  $M(r \Box \rightarrow s)$ , or  $M(r \rightarrow 3 s)$ , or  $r \Box \rightarrow Ms$ , or  $r \rightarrow 3 Ms$ ? None of these are adequate.

The problem is that (3) doesn't say, of any particular proposition, that it is true at most times. It doesn't say that among all times, most are such-and-such. Rather, it says that *among times at which it rains*, most times are times at which we stay inside. The function of the 'if'-clause in (3) is to **restrict the domain** of times over which the 'usually' operator quantifies.

Now return to (1) and (2). Suppose that here, too, the 'if'-clause serves to restrict the domain of times, so that 'always' and 'sometimes' only quantify over times at which it rains. On that hypothesis, (1) says that *among times at which it rains*, all times are times at which we stay inside, and (2) says that *among times at which it rains*, some times are times at which we stay inside. This is indeed what (1) and (2) mean, on their most salient interpretation.

As it turns out, 'among  $r$ -times, all times are  $s$ -times' is equivalent to 'all times are not- $r$ -times or  $s$ -times'. That's why we can formalize (1) as  $\Box(r \rightarrow s)$ . 'Among  $r$ -times, some times are  $s$ -times', on the other hand, is equivalent to 'some times are  $r$ -times and  $s$ -times'. That's why we can formalize (2) as  $\Diamond(r \wedge s)$ . It would be wrong to think that the conditional in (1) is material, the conditional in (2) is a conjunction, and the conditional in (3) is something else altogether. A much better explanation is that the 'if'-clause in (1) does the exact same thing as in (2) and (3). In each case, it restricts the domain of times over which the relevant operators quantify.

We can arguably see the same effect in (4) and (5).

(4) If the lights are on, Ada must be in her office.

- (5) If the lights are on, Ada might be in her office.

Letting the box express epistemic necessity, we can translate (4) as  $\Box(p \rightarrow q)$ . But (5) can't be translated as  $\Diamond(p \rightarrow q)$ , which would be equivalent to  $\Diamond(\neg p \vee q)$ . Nor can we translate (5) as  $p \rightarrow \Diamond q$ , which is entailed by  $\Diamond q$ . It is easy to think of scenarios in which (5) is false even though 'Ada might be in her office' is true. The correct translation of (5) is plausibly  $\Diamond(p \wedge q)$ . The sentence is true iff there is an epistemically accessible world at which the lights are on and Ada is in her office.

As before, we can understand what is going if we assume that the 'if'-clause in (4) and (5) functions as a restrictor. The 'if'-clause restricts the domain of worlds over which 'must' and 'might' quantify. (4) says that *among epistemically possible worlds at which the lights are on*, all worlds are worlds at which Ada is in her office. (5) says that *among epistemically possible worlds at which the lights are on*, some worlds are worlds at which Ada is in her office.

#### Exercise 8.16

Translate 'all dogs are barking' and 'some dogs are barking' into the language of predicate logic. Can you translate 'most dogs are barking' if you add a 'most' quantifier  $M$  so that  $Mx Fx$  is true iff most things satisfy  $Fx$ ?

The hypothesis that 'if'-clauses are restrictors also sheds light on the problem of conditional obligation.

- (6) Jones ought to help his neighbours.  
 (7) If Jones doesn't help his neighbours, he ought to not tell them that he's coming.

In chapter 6, we analyzed 'ought' as a quantifier over the best of the circumstantially accessible worlds. On this approach, (6) says that among the accessible worlds, all the best ones are worlds at which Jones helps his neighbours. Suppose the 'if'-clause in (7) serves to restrict the domain of worlds, excluding worlds at which Jones helps his neighbours. We then predict (7) to state that *among the accessible worlds at which Jones doesn't help his neighbours*, all the best worlds are worlds at which Jones doesn't tell his neighbours that he's coming. This can't be expressed by combining the monadic  $O$  quantifier with truth-functional connectives. Hence we had to introduce a primitive binary operator  $O(\cdot/\cdot)$ .

The upshot of all this is that we can make sense of a wide range of puzzling phenomena by assuming that 'if'-clauses are restrictors. Their function is to restrict the domain or worlds or times over which modal operators quantify.

What, then, is the purpose of ‘if’-clauses in “bare” conditionals like (8) and (9), where there are no modal operators to restrict?

(8) If Shakespeare didn’t write *Hamlet*, then someone else did.

(9) If Shakespeare hadn’t written *Hamlet*, then someone else would have.

Here opinions vary. One possibility, prominently defended by the linguist Angelika Kratzer, is that even bare conditionals contain modal operators. Arguably, ‘would’ in (9) functions as a kind of box. If this box is a simple quantifier over circumstantially accessible worlds, and the ‘if’-clause in (9) restricts its domain, then (9) can be formalized as  $\Box(p \rightarrow q)$ . If, on the other hand, ‘would’ in (9) works more like ‘ought’ – if it quantifies over the *closest* of the accessible worlds –, and the ‘if’-clause restricts the domain of accessible worlds, then the resulting truth-conditions are those of  $p \Box \rightarrow q$ . Both the strict and the variably strict analysis of (9) are therefore compatible with the hypothesis that ‘if’-clauses are restrictors.

What about (8)? This sentence really doesn’t appear to contain a relevant modal. Kratzer suggests that it contains an unpronounced epistemic ‘must’: (8) says that if Shakespeare didn’t write *Hamlet* then someone else *must* have written *Hamlet*. Assuming that the ‘if’-clause restricts the domain of this operator, bare indicative conditionals would be equivalent to strict epistemic conditionals.

### Exercise 8.17

Suppose bare indicative conditionals like (8) contain a box operator  $\Box$  whose accessibility relation relates each world to itself and to no other world. (This is a redundant operator insofar as  $\Box A$  is equivalent to  $A$ .) Assume the ‘if’-clause restricts the domain of that operator. What are the resulting truth-conditions of (8)?

### Exercise 8.18

Besides “would counterfactuals” there are also “might counterfactuals” like

(10) If I had played the lottery, I might have won.

Suppose ‘might’ is the dual of ‘would’, and suppose the ‘if’-clause in (10) restricts the domain of worlds over which ‘might’ quantifies. It follows that ‘if  $A$  then might  $B$ ’ is true iff  $B$  holds at some of the closest/accessible  $A$ -worlds. (‘Closest’ or ‘accessible’ depending on how we understand the ‘would’/‘might’ operators.) Can you see why this casts doubt on the validity of Conditional Excluded Middle?

## 9 Towards Modal Predicate Logic

### 9.1 Predicate logic recap

In these last two chapters, we are going to add the resources of first-order predicate logic to those of propositional modal logic. Let's begin by reviewing the syntax and semantics of classical, non-modal predicate logic.

The language  $\mathcal{L}_P$  of first-order predicate logic consists of *predicates*  $F^0, F^1, F^2, \dots, G^0, G^1, G^2, \dots$ , *individual constants* (or *names*)  $a, b, c, \dots$ , *individual variables*  $x, y, z, \dots$ , 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 of  $\mathcal{L}_P$  are formed by conjoining a predicate with zero or more terms. Each predicate takes a fixed number of terms, as indicated by its numerical superscript:  $F^1$  is a *one-place* predicate that combines with one term to form a sentence,  $F^2$  is *two-place*, and so on. In practice, we usually omit the superscripts, because context makes clear what kind of predicate is in play.  $Fa \vee Gab$ , for example, is well-formed only if  $F$  is one-place and  $G$  two-place.

In English, a predicate is what is what you get when you remove all names from a sentence. Removing 'Bob' from 'Bob is hungry' yields the predicate '– is hungry'. From 'Bob is in Rome', we get the two-place predicate '– is in –'. From 'Bob saw Carol's father in Jerusalem', we could get the three-place-predicate '– saw –'s father in –'. When we translate from English, we normally translate English names into  $\mathcal{L}_P$ -names and (logically simple) English predicates into  $\mathcal{L}_P$ -predicates. 'Bob is in Rome' might become  $Fab$ , where  $a$  translates 'Bob',  $b$  'Rome', and  $F$  '– is in –'.

From atomic sentences, complex sentences are formed in the usual way by means of the truth-functional operators  $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$ .

Another way to construct a complex sentence from a simpler sentence is to add a quantifier in front of the simpler sentence. A *quantifier* is an expression of the form  $\forall \chi$  or  $\exists \chi$ , where  $\chi$  is some variable. A quantifier is said to *bind* the variable it contains:  $\forall x$  binds  $x$ ,  $\exists y$  binds  $y$ , and so on.

In English, quantifier expressions are usually restricted to a particular subclass of the

things under discussion: ‘*all whales* are mammals’, ‘*some students* went home’. The  $\mathcal{L}_P$ -quantifiers  $\forall x$  and  $\exists x$  are unrestricted. They roughly correspond to ‘everything is such that ...’ and ‘something is such that ...’. We can translate restricted quantifiers by combining unrestricted quantifiers with truth-functional connectives. ‘All whales are mammals’ is equivalent to ‘Everything is either not a whale or a mammal’; so it can be translated as  $\forall x(Wx \rightarrow Mx)$ . ‘Some students went home’ could be translated as  $\exists x(Sx \wedge Hx)$ .

Variables are book-keeping devices. They function somewhat like pronouns in English.  $\exists x(Sx \wedge Hx)$  might be read as ‘something is such that *it* is a student and *it* went home’. By using different variables ( $x, y, z, \dots$ ), we can disambiguate statements with nested quantifiers. Consider

Every dog barked at a tree.

This can mean that there is a particular tree at which all the dogs barked, but it can also mean that each dog found some tree to bark at – possibly different trees for different dogs. The first reading could be translated as

$$\exists y(Ty \wedge \forall x(Dx \rightarrow Bxy)),$$

the second as

$$\forall x(Dx \rightarrow \exists y(Ty \wedge Bxy)).$$

Some more terminology. Recall that the *scope* of an operator (token) in a sentence is the shortest well-formed subsentence in which it occurs. In  $\exists y(Ty \wedge \forall x(Fx \rightarrow Bxy))$ , the scope of the quantifier  $\forall x$  is the subsentence  $\forall x(Fx \rightarrow Bxy)$ . If an occurrence of a variable lies in the scope of a quantifier that binds the variable, then the occurrence is called *bound*, otherwise it is *free*. In  $\forall x(Fx \rightarrow Bxy)$ , all occurrences of  $x$  are bound, but  $y$  is free.

A sentence containing free variables is called *open*. Sentences that aren’t open are *closed*. Intuitively, only closed sentences make complete statements. For this reason, some authors reserve the word ‘sentence’ for closed sentences, referring to open sentences as ‘formulas’. (Others call every  $\mathcal{L}_P$ -sentence a ‘formula’.)

### Exercise 9.1

Translate the following sentences into  $\mathcal{L}_P$ .

- (a) Keren and Keziah are sisters of Jemima.

- (b) All myriapods are oviparous.
- (c) Fred has a new car.
- (d) Not every student loves logic.
- (e) Every student who loves logic loves something.

Like sentences of modal propositional logic, sentences of predicate logic are interpreted 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. If we read  $\forall x$  as ‘everything is such that’ and  $\exists x$  as ‘something is such that’ then the relevant “somethings” are the members of the domain  $D$ .

The remainder of a model is an *interpretation function*  $V$  that assigns

- (a) to each name a member of  $D$ ,
- (b) to each zero-place predicate a truth-value,
- (c) to each one-place predicate a subset of  $D$ , and
- (d) to each  $n$ -place predicate with  $n > 1$  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 Bob is a member of  $D$ , then  $\langle \text{Bob}, \text{Bob} \rangle$  counts as a 2-tuple from  $D$ . (2-tuples are more commonly called *pairs*.) We can subsume condition (c) under condition (d) by assuming that a 1-tuple from  $D$  is a member of  $D$ . We can subsume (b) under (d) by identifying the truth-value False with the empty tuple  $\emptyset$  and the truth-value True with  $\{\emptyset\}$ . (Don’t worry if you find this confusing or objectionable. We won’t be using zero-ary predicates.)

### Definition 9.1

A **(classical) first-order model** is a pair  $\langle D, V \rangle$  consisting of

- a non-empty set  $D$ , and
- a function  $V$  that assigns to each name a member of  $D$  and to each  $n$ -place predicate a set of  $n$ -tuples from  $D$ .

As always, the purpose of a model is to represent a conceivable scenario together with an interpretation of the non-logical vocabulary. The non-logical vocabulary of  $\mathcal{L}_P$  are the names and predicates, which is why these are interpreted by  $V$ .

We assume that in any relevant scenario there are some things we want to talk about; these things are represented by the domain. The members of  $D$  are often called *individ-*

*uals*, but this should not be taken to imply anything about their nature. An individual might be a rock, a person, a symphony, a sentence, a number, or a possible world. Every  $\mathcal{L}_P$ -name is assumed to pick out one of these individuals. (Different names can pick out the same individual, and there can be individuals that aren't picked out by any name.)

Intuitively, a predicate expresses a property or relation that may be instantiated by the individuals in the domain. In order to determine the truth-value of a sentence like  $Fa$  or  $\exists xFx$  in a given scenario, however, we only need to know which individuals in the domain have the property expressed by  $F$ . Similarly, to determine the truth-value of sentences like  $Rab$  or  $\forall x\exists yRxy$ , we only need to know which pairs of individuals stand in the relation expressed by  $R$ . That's why the interpretation function in a first-order model 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  $V(a) \in V(F)$ . Likewise,  $Rab$  is true in a model iff the pair of individuals assigned to  $a$  and  $b$  – the pair  $\langle V(a), V(b) \rangle$  – is in the set assigned to  $R$ .

In this way, the truth-value of every closed atomic sentences is determined. For truth-functionally complex sentences, the standard rules apply: a negated sentence  $\neg A$  is true iff the corresponding sentence  $A$  is not true;  $A \wedge B$  is true iff  $A$  and  $B$  are both true; and so on.

When we turn to quantified sentences, we face a problem. We can't define the truth-value of  $\forall xFx$  in terms of the truth-value of  $Fx$ , because an open sentence like  $Fx$  doesn't have a truth-value. Interpretation functions interpret names and predicates; they say nothing about variables. Even if we changed this and said that  $x$  should also be interpreted as picking out a member of the domain, we would have to ignore this interpretation if we evaluate  $\forall xFx$ . 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, Alice and Bob, which are picked out by the names  $a$  and  $b$  respectively. Let  $V(F)$  be the set  $\{ \text{Alice} \}$ , a set that only contains Alice. So  $Fa$  is true and  $Fb$  false. The sentence  $Fx$  is neither true nor false, for the variable  $x$  does not refer to any particular individual. All we can say is that  $Fx$  is “true of” Alice and “false of” Bob. That is,  $Fx$  is true if we assign Alice to  $x$  and false if we assign Bob to  $x$ .  $\exists xFx$  is true because there is an individual (Alice) of which  $Fx$  is true. Equivalently,  $\exists xFx$  is true because there is some assignment of individuals to variables relative to which  $Fx$  is true.  $\forall xFx$  is false because it is not the case that every assignment of individuals to variables renders  $Fx$  true.

So we'll define truth relative to a model  $M = \langle D, V \rangle$  and a variable assignment  $g$ . A *variable assignment* is a function that maps variables to members of  $D$ . If we have

nested quantifiers, as in  $\forall x \exists y Gxy$ , we need to consider variable assignments that differ from other assignments with respect to a particular variable.  $\forall x \exists y Gxy$  is true iff, no matter what individual is assigned to  $x$ , there is some assignment of an individual to  $y$  (but holding fixed the assignment to  $x$ ) that makes  $Gxy$  true. Equivalently:  $\forall x \exists y Gxy$  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  $y$  such that  $Gxy$  is true relative to  $g'$ .

Let's say that (for any variable  $\chi$ ) a variable assignment  $g'$  is an  $\chi$ -variant of a variable assignment  $g$  iff  $g'$  differs from  $g$  at most in the value it assigns to  $\chi$ . Let's also introduce  $[\tau]^{M,g}$  as shorthand for the individual picked out by a term  $\tau$  in a model  $M = \langle D, V \rangle$  relative to assignment  $g$ :

$$[\tau]^{M,g} =_{\text{def}} \begin{cases} V(\tau) & \text{if } \tau \text{ is a name} \\ g(\tau) & \text{if } \tau \text{ is a variable.} \end{cases}$$

This is a compact way of saying that (1) for any variable  $\chi$ ,  $[\chi]^{M,g}$  is the individual assigned to  $\chi$  by  $g$ , and (2) for any name  $\eta$ ,  $[\eta]^{M,g}$  is the individual assigned to  $\eta$  by the interpretation function of  $M$ .

Now we can state the standard semantics of first-order predicate logic. (' $M, g \models A$ ' is pronounced ' $A$  is true in  $M$  relative to  $g$ ').

**Definition 9.2: Semantics of first-order predicate logic**

If  $M = \langle D, V \rangle$  is a first-order model,  $\phi^n$  is an  $n$ -place predicate (for  $n \geq 0$ ),  $\tau_1, \dots, \tau_n$  are terms,  $\chi$  is a variable, and  $g$  is a variable assignment, then

- (a)  $M, g \models \phi^n \tau_1 \dots \tau_n$  iff  $\langle [\tau_1]^{M,g}, \dots, [\tau_n]^{M,g} \rangle \in V(\phi)$ .
- (b)  $M, g \models \neg A$  iff  $M, g \not\models A$ .
- (c)  $M, g \models A \wedge B$  iff  $M, g \models A$  and  $M, g \models B$ .
- (d)  $M, g \models A \vee B$  iff  $M, g \models A$  or  $M, g \models B$ .
- (e)  $M, g \models A \rightarrow B$  iff  $M, g \not\models A$  or  $M, g \models B$ .
- (f)  $M, g \models A \leftrightarrow B$  iff  $M, g \models A \rightarrow B$  and  $M, g \models B \rightarrow A$ .
- (g)  $M, g \models \forall \chi A$  iff  $M, g' \models A$  for all  $\chi$ -variants  $g'$  of  $g$ .
- (h)  $M, g \models \exists \chi A$  iff  $M, g' \models A$  for some  $\chi$ -variant  $g'$  of  $g$ .

Clause (a) says that, for example,  $Fa$  is true in a model  $M$  relative to an assignment  $g$  iff in that model, the predicate  $F$  applies to the individual picked out by  $a$ . Clauses (b)-(f) say that the truth-functional operators are interpreted in the standard fashion. Clauses

(g) and (h) tell us how quantified sentences are interpreted.  $\exists xFx$ , for example, is true relative to  $M$  and  $g$  iff  $Fx$  is true relative to some assignment function  $g'$  that differs from  $g$  at most in what it assigns to  $x$ .

Definition 9.2 settles the truth-value of every  $\mathcal{L}_P$ -sentence in every (first-order) model, relative to any assignment function.

We can also define a concept of truth relative to a model, without reference to an assignment function. Let's say that an  $\mathcal{L}_P$ -sentence is **true in a model**  $M$  iff it is true in  $M$  relative to *every* assignment function  $g$  for  $M$ .

Finally, we say that an  $\mathcal{L}_P$ -sentence is **valid** (in classical first-order logic) iff it is true in all (classical, first-order) models. Equivalently: An  $\mathcal{L}_P$ -sentence is valid iff it is true in all models relative to all assignment functions.

On the present definition,  $Fx \rightarrow Fx$  is valid, even though it does not make a complete statement, due to the free variable  $x$ . To avoid this, some authors restrict the concept of validity to closed sentences.

### Exercise 9.2

Define a first-order model in which  $\exists xFx \rightarrow \forall xFx$  is false. Demonstrate that the sentence is false in your model by applying all relevant clauses from definition 9.2.

### Exercise 9.3

The definition of truth in a model uses the method of supervaluation that we met in section 7.4. Give examples to illustrate the following claims.

- (a) If a sentence  $A$  is not true in a model, it does not follow that  $\neg A$  is true in the model.
- (b) A disjunction  $A \vee B$  can be true in a model even though neither  $A$  nor  $B$  is true in the model.

## 9.2 Modal fragments of predicate logic

Much of the power and complexity of predicate logic comes from its ability to handle nested quantifiers with different variables. For some applications, these complexities aren't needed, and we can simplify the semantics.

Consider a fragment  $\mathcal{L}_P^1$  of  $\mathcal{L}_P$  with only one variable  $x$ , no names, and only one-place predicates. In  $\mathcal{L}_P^1$ , we have sentences like  $Fx$ ,  $\forall xGx$ ,  $\forall x\exists x(Fx \rightarrow Gx)$ , but not  $Fa$  or  $\forall x\exists y(Fx \rightarrow Gy)$ .

Following definition 9.1, a model for  $\mathcal{L}_P^1$  consists of a non-empty set  $D$  and an interpretation function  $V$  that assigns to each predicate a subset of  $D$ . That is, for  $\mathcal{L}_P^1$  definition 9.1 can be simplified as follows:

A **model of  $\mathcal{L}_P^1$**  is a pair  $\langle D, V \rangle$  consisting of

- a non-empty set  $D$ , and
- a function  $V$  that assigns to every  $\mathcal{L}_P^1$ -predicate a subset of  $D$ .

We can also simplify definition 9.2. Since  $\mathcal{L}_P^1$  has only one variable  $x$ , an assignment function for  $\mathcal{L}_P^1$  only needs to tell us which individual in  $D$  is picked out by  $x$ . So we can represent an entire assignment function for  $\mathcal{L}_P^1$  by a member of  $D$ . This leaves us with the following semantics.

If  $M = \langle D, V \rangle$  is a model for  $\mathcal{L}_P^1$ ,  $d$  is a member of  $D$ , and  $\phi$  is an  $\mathcal{L}_P^1$ -predicate, then

- (a)  $M, d \models \phi x$       iff  $d \in V(\phi)$ .
- (b)  $M, d \models \neg A$       iff  $M, d \not\models A$ .
- (c)  $M, d \models A \wedge B$     iff  $M, d \models A$  and  $M, d \models B$ .
- (d)  $M, d \models A \vee B$     iff  $M, d \models A$  or  $M, d \models B$ .
- (e)  $M, d \models A \rightarrow B$     iff  $M, d \not\models A$  or  $M, d \models B$ .
- (f)  $M, d \models A \leftrightarrow B$     iff  $M, d \models A \rightarrow B$  and  $M, d \models B \rightarrow A$ .
- (g)  $M, d \models \forall x A$       iff  $M, d' \models A$  for all  $d' \in D$ .
- (h)  $M, d \models \exists x A$       iff  $M, d' \models A$  for some  $d' \in D$ .

These definitions look a lot like definitions 2.1 and 2.2 from chapter 2. The only difference is that the sentence letters from chapter 2 are now called predicates and written in uppercase, the box is written  $\forall x$ , the diamond  $\exists x$ , and we always append the letter  $x$  to sentence letters: we write  $\forall x Fx$ , not  $\forall x F$ . But it doesn't really matter how a symbol is called or how it is written.

The upshot is that propositional modal logic, interpreted as in chapter 2, can be regarded as a disguised *fragment of first-order predicate logic*. The sentence letters of  $\mathcal{L}_M$  are disguised (one-place) predicates, the box and the diamond are disguised quantifiers. If we adopted the orthographic convention to write the box as  $\forall x$ , the diamond as  $\exists x$ , and to always append the letter  $x$  to (capitalised) sentence letters,  $\mathcal{L}_M$  would look just

like  $\mathcal{L}_P^1$ , and it would have the same semantics.

If we use chapter 3's Kripke semantics rather than the simple semantics from chapter 2 to interpret  $\mathcal{L}_M$ , we get a different fragment of first-order predicate logic. The box and the diamond are still disguised quantifiers, but this time they are restricted by the accessibility relation. We could drop the disguise by writing  $\Box p$  as  $\forall y(Rxy \rightarrow Py)$  and  $\Diamond p$  as  $\exists y(Rxy \wedge Py)$ . The fragment of  $\mathcal{L}_P$  that now corresponds to  $\mathcal{L}_M$ -sentences has two variables  $x$  and  $y$  and one two-place predicate ' $R$ ' in addition to the one-place predicates; it no longer has unrestricted quantifiers.

What's the point of the disguise? Why didn't we write boxes and diamonds as  $\mathcal{L}_P$ -quantifiers all along? There are several reasons.

One is that we often use the box and the diamond to formalize pre-theoretic concepts of which it is not obvious that they can be understood as a quantifiers over worlds. Some hold that the correct semantics for obligation and permission, for example, is not Kripke semantics, but neighbourhood semantics. The language of modal propositional logic is neutral on this disagreement. Or think of provability logic, where the box formalizes mathematical provability. As it turns out, one can give a Kripke semantics for provability, but nobody thinks that this somehow reveals what provability really means. In provability logic,  $\Box A$  means that  $A$  is derivable from the axioms and rules of (say) ZFC; it would not be illuminating to write this as  $\forall y(Rxy \rightarrow Ay)$ .

One might also argue that the syntax of modal logic conveniently resembles the surface form of English statements that we may want to formalize. In 'Bob knows that it is raining', for example, the object of Bob's knowledge is specified by 'it is raining'. It seems appropriate to formalize the sentence in terms of an operator  $K$  that applies to a sentence,  $p$ . If we "dropped the disguise", the formalization would be  $\forall y(Rxy \rightarrow Py)$ . The sentence 'it is raining' would have to be translated by a predicate  $P$  – a predicate that applies to all and only the worlds at which it is raining.

There is a deeper point here. Sentences of modal logic are interpreted *at a world* in a model. Modal logic looks at models "from the inside", from the perspective of a particular world. Predicate logic, by contrast, describes models "from the outside", from a God's eye perspective. If we want to say that a particular individual has a property  $P$  in predicate logic, we need to pick out that individual among all the elements of the domain, perhaps by a name. We can then say  $Pa$ . In modal logic, we can simply say  $p$  to express that the internal point from which we're looking at the model has the relevant property.

For many applications, this internal perspective is natural. If we think about what is possible or what the future will bring, our thinking takes place at a particular time, in a particular world. We are looking at the structure of times and worlds from the inside. When I say that it is raining, I mean that it is raining *here and now*, in *this world*. I don't

need to pick out the relevant time and place and world from a God’s eye perspective. I can pick them out simply as the time and place and world at which I currently find myself.

There are other, more pragmatic reasons to use the modal language  $\mathcal{L}_M$  rather than  $\mathcal{L}_P$ . The language of boxes and diamonds is simpler than the language of first-order predicate logic. It has a simpler syntax, a simpler semantics, and allows for simpler proofs. For almost all the conceptions of validity we have studied (K-validity, S4-validity, etc.), there are efficient mechanical procedures to determine whether an arbitrary  $\mathcal{L}_M$ -sentence is valid or invalid. By contrast, there is no mechanical procedure at all to determine, for an arbitrary  $\mathcal{L}_P$ -sentence, whether it is valid or invalid.

You may wonder how this is possible given that  $\mathcal{L}_M$ -sentences are just  $\mathcal{L}_P$ -sentences in disguise. The reason is that while every  $\mathcal{L}_M$ -sentence is a disguised  $\mathcal{L}_P$ -sentence, not every  $\mathcal{L}_P$ -sentence can be disguised as an  $\mathcal{L}_M$ -sentence. There are many things one can say in  $\mathcal{L}_P$  that can’t be said in  $\mathcal{L}_M$ . The  $\mathcal{L}_P$ -sentence  $\forall x Rxx$ , for example, states that  $R$  is reflexive. No sentence of  $\mathcal{L}_M$  has this meaning: there is no  $\mathcal{L}_M$ -sentence that is true at a world in a model iff the model’s accessibility relation is reflexive.

That’s why modal propositional logic, interpreted as in chapter 2 or 3, is a disguised *fragment* of predicate logic. It is a simple and computationally attractive fragment that takes an “internal” perspective on models.

#### Exercise 9.4

Since  $\Box A \rightarrow A$  corresponds to reflexivity, one might think that  $\Box p \rightarrow p$  is true at a world in a model iff the model’s accessibility relation is reflexive. (a) Explain why this is not correct. (b) Can you show that there is no  $\mathcal{L}_M$ -sentence that is true at a world in a model iff the model’s accessibility is reflexive?

### 9.3 Predicate logic proofs

If we want to know whether an  $\mathcal{L}_P$ -sentence is valid or invalid, we could in principle work through definition 9.2. Various proof systems for classical predicate logic offer a more streamlined approach.

Let’s look at the tree method for classical predicate logic. Suppose we want to test whether  $\exists x(Fx \wedge Gx) \rightarrow \exists xFx$  is valid. As always, we start the tree with the negation of the target sentence:

$$1. \neg(\exists x(Fx \wedge Gx) \rightarrow \exists xFx) \quad (\text{Ass.})$$

There is no world label because we're not doing modal logic. Next, we apply the standard rule for negated conditionals:

2.  $\exists x(Fx \wedge Gx)$  (1)
3.  $\neg \exists xFx$  (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)

We expand the conjunction on node 4.

5.  $Fa$  (4)
6.  $Ga$  (4)

Next, we expand node 3, which says that  $Fx$  is true of nothing. In particular then,  $Fx$  can't be true of  $a$ . So we add  $\neg Fa$ :

7.  $\neg Fa$  (3)
- ×

The tree is closed because the sentence on node 7 is the negation of the sentence on node 5. The target sentence is valid.

To state the general rules, we need some more notation. If  $A$  is a sentence and  $\tau_1, \tau_2$  terms, let  $A[\tau_2/\tau_1]$  be the sentence obtained from  $A$  by replacing all free occurrences of  $\tau_1$  with  $\tau_2$ . 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  $\exists \chi A$  is that you add a node  $A[\eta/\chi]$ , where  $\eta$  is a “new” name that does not already occur on the relevant branch. If such a node has been added to every open branch below  $\exists \chi A$  then the  $\exists \chi A$  node can be ticked off.  $\forall \chi A$  nodes can be expanded multiple times, once for each “old” name. So if  $\forall xA$  occurs on a branch, and the branch contains the names  $a$  and  $b$  then we can add both  $A[a/x]$  and  $A[b/x]$ . If there is no old name on a branch, we are allowed to expand  $\forall \chi A$  with a new name.  $\forall \chi A$  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.

$\forall \chi A$	$\exists \chi A$	$\neg \forall \chi A$	$\neg \exists \chi A$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$A[\eta/\chi]$	$A[\eta/\chi]$	$\neg A[\eta/\chi]$	$\neg A[\eta/\chi]$
$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
old or first	new	new	old or first

If you want to read off a countermodel from an open branch, you can simply take the domain  $D$  to consist of all names that occur on the branch. For the interpretation function  $V$ , you then stipulate that each name picks out itself – so that, for example,  $V(a) = a$  – and that a predicates  $P$  applies to a tuple of names  $\langle a, b, \dots \rangle$  iff  $Pab \dots$  occurs on the branch.

### Exercise 9.5

Give tree proofs for the following sentences.

- (a)  $\forall x Fx \rightarrow Fa$
- (b)  $\forall x (Fx \rightarrow Gx) \rightarrow (\forall x Fx \rightarrow \forall x Gx)$
- (c)  $\forall x (Fx \wedge Gx) \leftrightarrow (\forall x Fx \wedge \forall x Gx)$
- (d)  $\exists x \forall y Gxy \rightarrow \forall y \exists x Gxy$
- (e)  $\exists y \forall x (Fy \rightarrow Fx)$

There are also axiomatic calculi for predicate logic. We can, for example, use the following axiom schemas:

- ( $\forall\exists$ )  $\neg \exists \chi A \leftrightarrow \forall \chi \neg A$
- (UI)  $\forall \chi A \rightarrow A[\eta/\chi]$
- (DI)  $\forall \chi (A \rightarrow B) \rightarrow (A \rightarrow \forall \chi B)$ , if  $\chi$  is not free in  $A$

To these we would add the following rules. As in earlier chapters,  $\Gamma \models_0 A$  means that  $A$  is a truth-functional consequence of (the sentences in)  $\Gamma$ .

- (CPL) If  $\Gamma \models_0 A$  and all members of  $\Gamma$  are on a proof, then one may add  $A$ .
- (Gen) If  $A$  occurs on a proof, then one may add  $\forall \chi A[\chi/\eta]$ .

These axioms and rules are sound and complete: everything that can be proved is valid, and every valid (closed) sentence can be proved. The above tree rules are also sound and complete.

### Exercise 9.6

The completeness proof for first-order trees (like the proof in chapter 4) shows that if a sentence is valid then any fully expanded tree for that sentence will close, provided the tree rules are applied in a sensible order. Why doesn't this contradict the claim I made in the previous section. that there is no mechanical procedure to determine, for an arbitrary  $\mathcal{L}_P$ -sentence, whether the sentence is valid? (Tree proofs count as "mechanical", so that's not the problem.)

## 9.4 Modality de dicto and de re

We are now ready to add boxes and diamonds to the language of first-order predicate logic. This gives us the **standard language of first-order modal logic**, or  $\mathcal{L}_{MP}$ . The sentences of  $\mathcal{L}_{MP}$  are defined as follows.

1. An  $n$ -place predicate followed by  $n$  terms is an  $\mathcal{L}_{MP}$ -sentence.
2. If  $A$  is an  $\mathcal{L}_{MP}$ -sentence, then so are  $\neg A$ ,  $\Diamond A$ , and  $\Box A$ .
3. If  $A$  and  $B$  are  $\mathcal{L}_{MP}$ -sentences, then so are  $(A \wedge B)$ ,  $(A \vee B)$ ,  $(A \rightarrow B)$  and  $(A \leftrightarrow B)$ .
4. If  $A$  is an  $\mathcal{L}_{MP}$ -sentence and  $\chi$  is a variable, then  $\forall \chi A$  and  $\exists \chi A$  are  $\mathcal{L}_{MP}$ -sentence.
5. Nothing else is an  $\mathcal{L}_{MP}$ -sentence.

We continue to interpret the box and the diamond as (disguised) quantifiers. So  $\mathcal{L}_{MP}$  effectively has two kinds of quantifiers: overt quantifiers of the form  $\forall \chi$  and  $\exists \chi$ , and the disguised quantifiers  $\Box$  and  $\Diamond$ . This is only useful if the two kinds of quantifiers range over different things. In applications of modal predicate logic, the box and the diamond usually range over possible worlds or times, while the overt quantifiers range over things like people, rocks, ghosts, etc., which are assumed to inhabit the worlds or times.

For example, consider the following inference, in which I've written the box as 'K'.

Bob knows that all humans are mortal.	$K \forall x(Hx \rightarrow Mx)$
Socrates is human.	$HS$
Therefore: Socrates is mortal.	$MS$

The knowledge operator  $K$  is a quantifier over the worlds compatible with Bob's (implicit) knowledge.  $K \forall x(Hx \rightarrow Mx)$  says that  $\forall x(Hx \rightarrow Mx)$  is true at every world compatible with Bob's knowledge.  $\forall x(Hx \rightarrow Mx)$  is assumed to quantify not over worlds, but over things that exist relative to a world.  $\forall x(Hx \rightarrow Mx)$  is true at a world  $w$  iff  $Hx \rightarrow Mx$  is true

of every inhabitant of  $w$ , meaning that every inhabitant of  $w$  is either not human or mortal. The inference is valid because the accessibility relation for knowledge is reflexive.

Imagine a lottery. Let's read the box as 'it is certain that' and  $W$  as '– is a winning ticket'. Can you see what is expressed by the following two statements?

- (1)  $\Box \exists x Wx$
- (2)  $\exists x \Box Wx$

(1) says that it is certain that some ticket wins: at every epistemically accessible world there is a winning ticket. (2) says that there is a particular ticket of which we are sure that it will win: there is an individual such that at every epistemically accessible world, *it* is the winning ticket. (2) is only true if we know which ticket is the (or a) winning ticket.

Sentences like  $\exists x \Box Wx$  are called **de re**, Latin for 'of a thing'. Intuitively,  $\exists x \Box Wx$  assert *of* a particular ticket that it has a modal property, namely the property of being the certain winner. By contrast,  $\Box \exists x Fx$ , merely states that the proposition (Latin, *dictum*)  $\exists x Fx$  is certain. Sentences like this are called **de dicto**.

In general, an  $\mathcal{L}_{MP}$ -sentence is *de re* whenever it contains a variable that is free in the scope of some modal operator. To determine whether a sentence  $A$  is *de re*, first identify all subsentences of  $A$  that constitute the scope of a modal operator. (In  $\exists x \Box Wx$ , there is one such subsentence:  $\Box Wx$ .) Next, check if at least one of these subsentences contains a free variable. ( $\Box Wx$  contains the free variable  $x$ .) If yes, the sentence  $A$  is *de re*.

If a sentence contains a modal operator and is not *de re*, then it is *de dicto*. So  $\forall x(Fx \rightarrow \Box Gx)$  and  $\exists y \Box (\forall x Fx \rightarrow Fy)$  are *de re*, but  $\Box \forall x Fx \rightarrow Fa$  is *de dicto*.  $\forall x Fx \rightarrow Fa$  is neither *de dicto* nor *de re*, because it isn't modal.

There is no consensus on how to classify sentences like  $\Box Fa$  that contain a name, but no free variable, in the scope of a modal operator. One might argue that  $\Box Fa$  is *de dicto* because it attributes a modal status – say, necessity – to the proposition  $Fa$ . But one might also interpret the sentence as attributing a modal property to the individual  $a$ : the property of being necessarily  $F$ . The sentence should then be classified as *de re*. Which of these two perspectives is more adequate depends on the precise semantics of  $\mathcal{L}_{MP}$ . We therefore have to postpone the question until the next chapter, where we will consider some options for developing a semantics of  $\mathcal{L}_{MP}$ .

Many natural-language sentences are ambiguous between a *de re* reading and a *de dicto* reading. Consider 'something necessarily exists'. This can mean either that there is an object which could not have failed to exist ( $\exists x \Box Ex$ ); but it can also mean that it is necessary that something or other exists ( $\Box \exists x Ex$ ). The first reading is *de re*, the second *de dicto*.

### Exercise 9.7

Translate the following sentences into modal predicate logic. (Some of them are ambiguous.)

- (a) John must be hungry.
- (b) Anyone who is a cyclist must have legs.
- (c) Every day might be our last.
- (d) If anyone wants to leave early, they should do so quietly.
- (e) Everyone who bought a ticket is allowed to enter.

### Exercise 9.8

Which of your translations from the previous exercise are *de re* and which are *de dicto*?

On some interpretations of the modal operators, one may question whether *de re* sentences are intelligible. Suppose we interpret the box as ‘it is analytic that’ or ‘it is provable that’. The things that are analytic or provable are sentences or propositions.  $2+2=4$ , for example, is provable in ZFC, and ‘all vixens are female foxes’ is analytic in English. (Remember that a sentence is analytic if it is true in virtue of its meaning.) It is not clear what it could mean to say that something is provable or analytic *of* a particular thing.

To illustrate the problem, let’s introduce the name ‘Julius’ for whoever invented the zip. The sentence ‘Julius invented the zip’ is analytic. (In fact, ‘Julius invented the zip’ entails that someone invented the zip, which is not analytic. We should really use ‘If anyone invented the zip, then Julius invented the zip’. Let’s ignore this complication.) But is it 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 his or her parents, the sentence is not analytic.

This kind of worry was prominently raised by W.V.O. Quine in the 1940s. It has since faded, mostly because philosophers have turned their attention away from analyticity to other interpretations of the box for which the problem is thought not to arise. But we will return to the matter in section 10.4.

## 9.5 Identity and descriptions

In applications of modal and non-modal predicate logic, it is often useful to have a special predicate for identity. Let's assume that  $\mathcal{L}_P$  and  $\mathcal{L}_{MP}$  have the two-place predicate '='. The identity predicate is conventionally placed between its two arguments: we write ' $a = b$ ', not ' $=ab$ '. We also sometimes abbreviate ' $\neg a = b$ ' as ' $a \neq b$ '.

Unlike the other predicates of  $\mathcal{L}_P$  and  $\mathcal{L}_{MP}$ , the identity predicate counts as a logical symbol. Its meaning is held fixed. In any model,  $a = b$  means that the individual picked out by  $a$  is the very same thing as the individual picked out by  $b$ . This is reflected by the following clause, which we add to the semantics of predicate logic:

$$M, g \models \tau_1 = \tau_2 \quad \text{iff} \quad [\tau_1]^{M, g} = [\tau_2]^{M, g}.$$

It is easy to see that the sentence  $a = a$  is now valid, because  $a$  and  $a$  are guaranteed to pick out the same individual. More interestingly, since the function of a name in classical predicate logic is just to pick out an individual, it never matters which of two names we use if they pick out the same individual. That is, if  $a = b$  is true, then replacing some or all occurrences of  $a$  in a sentence with  $b$  never affects whether the sentence is true. This principle is known as **Leibniz' Law**.

To reflect these facts, the tree method for (non-modal) predicate logic must be extended by two new rules. First, if  $\eta$  is an "old" name (that already occurs on a branch) then we can always add a node  $\eta = \eta$  to the branch. Second, if an identity statement  $\eta_1 = \eta_2$  occurs on a branch, and some sentence  $A$  on the branch contains  $\eta_1$ , then we may add a new node with the same sentence  $A$  except that one or more occurrences of  $\eta_1$  in  $A$  are replaced by  $\eta_2$ , or one or more occurrences of  $\eta_2$  by  $\eta_1$ . Let  $A[\eta_2//\eta_1]$  stand for any sentence that results from  $A$  by replacing one or more occurrences of  $\eta_1$  by  $\eta_2$ . The new rules can then be summarized as follows.

Self-Identity

$$\begin{array}{c} \vdots \\ \eta = \eta \\ \uparrow \\ \text{old} \end{array}$$

Leibniz' Law

$$\begin{array}{c} \eta_1 = \eta_2 \\ A \\ \vdots \\ A[\eta_2//\eta_1] \end{array}$$

Leibniz' Law

$$\begin{array}{c} \eta_1 = \eta_2 \\ A \\ \vdots \\ A[\eta_1//\eta_2] \end{array}$$

Here is a tree for  $(Raa \wedge a = b) \rightarrow Rab$ , using Leibniz's Law.

1.	$\neg((Raa \wedge a=b) \rightarrow Rab)$	(Ass.)
2.	$Raa \wedge a=b$	(1)
3.	$\neg Rab$	(1)
4.	$Raa$	(2)
5.	$a=b$	(2)
6.	$Rab$	(4, 5, LL)
	x	

### Exercise 9.9

Use the tree method to check which of the following sentences are valid.

- (a)  $\forall x(x=x)$
- (b)  $\forall x\forall y(x=y \rightarrow y=x)$
- (c)  $(a=b \wedge b=c) \rightarrow a=c$
- (d)  $Rab \rightarrow \forall x(x=a \leftrightarrow Rxb)$
- (e)  $\forall x\forall y\forall z((x \neq y \wedge y \neq z) \rightarrow x \neq z)$

### Exercise 9.10

Show that the second version of the Leibniz' Law rule is redundant: we could reach  $A[\eta_1//\eta_2]$  from  $\eta_1 = \eta_2$  and  $A$  with the other rules.

In the axiomatic approach, the two facts about identity are often represented by the following axiom schemas:

- (SI)  $\eta = \eta$
- (LL)  $\eta_1 = \eta_2 \rightarrow (A \rightarrow A[\eta_2//\eta_1])$

Once we add boxes and diamonds to the language of predicate logic, the seemingly harmless axioms and rules for identity become problematic. Consider the following inference:

It is analytic that Julius invented the zip.

Julius = Whitcomb L. Judson.

Therefore: It is analytic that Whitcomb L. Judson invented the zip.

The conclusion clearly doesn't follow from the premises, but the inference seems to be licensed by Leibniz's law. Another well-known example:

Lois Lane believes that Superman can fly.  
 Superman = Clark Kent.  
 Therefore: Lois Lane believes that Clark Kent can fly.

### Exercise 9.11

(a) Give an axiomatic proof of  $\Box \exists x x = a$ , using (SI), (UI), (CPL),  $(\forall \exists)$ , (CPL), and (Nec), in this order. (b) Can you see why we might not want to count  $\Box \exists x x = a$  as a logical truth in some applications of modal logic? At which point do you think the proof goes wrong?

We will return to these issues in section 10.4. In the remainder of the present section, I want to highlight some other things we can do with the identity predicate, apart from making claims about identity.

You have already encountered one other use in earlier chapters. Suppose we want to express that some relation  $R$  is connected, meaning that for any two things, either the first is  $R$ -related to the second or the second is  $R$ -related to the first. This can't be expressed without an identity predicate. With an identity predicate, it is easy:

$$\forall x \forall y (Rxy \vee x = y \vee Ryx).$$

We can also use identity to express numerical quantifiers. For example, we can express 'there are at least two  $F$ s' as

$$\exists x (Fx \wedge \exists y (Fy \wedge x \neq y)).$$

'There is exactly one  $F$ ' can be expressed as

$$\exists x (Fx \wedge \forall y (Fy \rightarrow x = y)).$$

### Exercise 9.12

Can you express the following in  $\mathcal{L}_P$  with identity?

- (a) There are exactly two  $F$ s.
- (b) There are no more than three  $F$ s.

Another important use of the identity predicate is to formalise statements involving definite descriptions. A definite description is a complex noun phrase, typically of the form 'the  $F$ ', that purports to pick out a particular object. 'The current Prime Minister',

‘the highest mountain in Scotland’, and ‘Carol’s father’ are definite descriptions.

The standard language of predicate logic does not have a definite article (‘the’). The only way to pick out an individual in  $\mathcal{L}_P$  is by a name. But there are good reasons not to translate descriptions as names.

One reason is that we would thereby miss logical connections between descriptions and predicates. ‘The current Prime Minister is not Prime Minister’ is a logical contradiction, but this can’t be brought out if we translate ‘the current Prime Minister’ as a simple name.

Another reason not to translate descriptions as names is that descriptions often give rise to a *de re/de dicto* ambiguity. Consider the following sentence:

The Pope might have been Italian.

This has two readings. It can mean either that the actual Pope, Jorge Mario Bergoglio, might have been Italian (*de re*). Alternatively, it can mean that the following might have been the case: some Italian person is Pope (*de dicto*). There is no way to account for these two readings in  $\mathcal{L}_{MP}$  if we translate ‘the Pope’ as a name.

A better translation for statements involving definite descriptions was proposed by Bertrand Russell in 1905. Russell argued that a statement of the form ‘the  $F$  is  $G$ ’ is true iff there is exactly one (relevant)  $F$ , and this one  $F$  is also  $G$ . If we have an identity predicate, we can easily express this in the language of predicate logic:

$$\exists x(Fx \wedge \forall y(Fy \rightarrow x=y) \wedge Gx).$$

Following Russell, we might translate ‘The current Prime Minister is not Prime Minister’ as

$$\exists x(Px \wedge \forall y(Py \rightarrow x=y) \wedge \neg Px).$$

This is indeed a contradiction: it is true in no model.

We can also account for the two readings of ‘the Pope might have been Italian’. The *de re* reading is

$$\exists x(Px \wedge \forall y(Py \rightarrow x=y) \wedge \Diamond Ix).$$

The *de dicto* reading is

$$\Diamond \exists x(Px \wedge \forall y(Py \rightarrow x=y) \wedge Ix).$$

**Exercise 9.13**

Give two translations for each of the following sentences, one *de re* and one *de dicto*.

- (a) Hillary Clinton might have been the 45th US President.
- (b) Smith's murderer could have been a woman.
- (c) Alice believes that the student representative is rude.



# 10 Semantics for Modal Predicate Logic

## 10.1 Constant domain semantics

We have met the language  $\mathcal{L}_{MP}$  of (first-order) modal predicate logic. It is time to think about how this language should be interpreted. This will tell us which sentences and inferences in the language are valid.

As in modal propositional logic, we will assume that the box and the diamond are quantifiers over accessible worlds, where “accessibility” is a placeholder whose meaning depends on the application. If we want to reason about knowledge, a world  $v$  would be accessible from a world  $w$  iff  $v$  is compatible with what is known at  $w$ . If we’re interested in metaphysical modality, a world  $v$  would be accessible from a world  $w$  iff it is compatible with the nature of things at  $w$ . And so on.

In a typical application, we use the names of  $\mathcal{L}_{MP}$  to pick out people or rocks or other things that might exist at a world. We use the predicates to express properties or relations. We might, for example, interpret  $\Diamond Fa$  as saying that Aristotle could have been a sailor, assuming that  $a$  picks out Aristotle and  $F$  the property of being a sailor.

Our concern in logic is not whether a particular claim about Aristotle is true. We want to know which statements are *logically true* or *valid*, meaning that they are true in any conceivable scenario, under any interpretation of the non-logical expressions (but holding fixed the meaning of the modal operators).

We use models to represent a scenario together with an interpretation of the non-logical vocabulary. A model for  $\mathcal{L}_{MP}$  contains just enough information about a scenario and an interpretation to determine, for every  $\mathcal{L}_{MP}$ -sentence and every world, whether the sentence is true at that world.

The non-logical vocabulary of  $\mathcal{L}_{MP}$  are the names and the predicates (with the exception of the identity predicate ‘=’). Let’s assume, for now, that the purpose of a name is simply to pick out an individual. Intuitively, a predicate picks out a property or relation. In non-modal predicate logic, we could represent these properties or relations by their extension – by the sets of individuals (or tuples of individuals) to which they apply. In modal predicate logic, we typically want to allow for scenarios in which an individual

has different properties at different worlds. In one world, Aristotle might be a sailor, in another he might be a shoemaker. If  $F$  expresses the property of being a sailor, then the set of individuals to whom  $F$  applies will differ from world to world. To determine the truth-value of  $Fa$  at a world, we need to know to which individuals  $F$  applies *at that world*. A model's interpretation function will therefore assign a set of (tuples of) individuals to each predicate *relative to each world*.

Consider a model with two worlds  $w$  and  $v$ . Both worlds, let's assume, are accessible from  $w$  and neither is accessible from  $v$ . The model's interpretation function tells us that the name  $a$  picks out Aristotle. It also tells us that the predicate  $F$  applies to Aristotle and Boethius at  $w$  and only to Boethius at  $v$ . We can write this as follows:

$$\begin{aligned} V(a) &= \text{Aristotle} \\ V(F, w) &= \{\text{Aristotle}, \text{Boethius}\} \\ V(F, v) &= \{\text{Boethius}\} \end{aligned}$$

We don't know what property is expressed by  $F$ , nor which properties Aristotle and Boethius have at  $w$  and  $v$ . Nonetheless, we can figure out that  $Fa$  is true at  $w$ , because the predicate  $F$  applies to Aristotle at  $w$ . We can also figure out that  $Fa$  is false at  $v$ , and that  $\Box Fa$  is false at  $w$ .

To determine the truth-value of arbitrary  $\mathcal{L}_{MP}$ -sentences, we need some more information. As it stands, we can't tell whether (say)  $\forall xFx$  is true at  $w$ . Informally,  $\forall xFx$  says that every individual is  $F$ . We know that Aristotle and Boethius are  $F$  at  $w$ . But we don't know if there are other individuals besides Aristotle and Boethius. If yes, then  $\forall xFx$  is false at  $w$ . If no, the sentence is true. We therefore assume that a model for  $\mathcal{L}_{MP}$  also specifies a domain of individuals.

### Definition 10.1

A **constant-domain Kripke model** for  $\mathcal{L}_{MP}$  is a structure  $M$  consisting of

1. a non-empty set  $W$  (the "worlds"),
2. a binary ("accessibility") relation  $R$  on  $W$ ,
3. a non-empty set  $D$  (of "individuals"), and
4. an interpretation function  $V$  that assigns
  - to each  $\mathcal{L}_{MP}$ -name a member of  $D$ , and
  - to each  $n$ -place predicate of  $\mathcal{L}_{MP}$  and world  $w \in W$  a set of  $n$ -tuples from  $D$ .

Models of this type are called “constant-domain models” because the domain of individuals is the same for each world. This may seem questionable – and we are soon going to question it – but it simplifies the semantics. Let’s stick with it for the moment.

Having defined a concept of a model, we can lay down the rules that determine whether any given  $\mathcal{L}_{MP}$ -sentence is true at a world in a model.

In fact, truth will be defined relative to three parameters: a model, a world, and an assignment function. The assignment function plays the same role as in non-modal predicate logic.  $\forall x \Diamond Fx$ , for example, is true at a world  $w$  in a model iff there is some assignment of an individual to  $x$  that renders  $\Diamond Fx$  true at  $w$ . We continue to use  $[\tau]^{M,g}$  for the individual picked out by a term (name or variable)  $\tau$  relative to a model  $M = \langle D, W, R, V \rangle$  and an assignment function  $g$ :

$$[\tau]^{M,g} =_{\text{def}} \begin{cases} V(\tau) & \text{if } \tau \text{ is a name} \\ g(\tau) & \text{if } \tau \text{ is a variable.} \end{cases}$$

### Definition 10.2: Constant-domain Kripke semantics

If  $M = \langle W, R, D, V \rangle$  is a constant-domain Kripke model,  $w$  is a member of  $W$ ,  $\phi$  is an  $n$ -place predicate (for  $n \geq 0$ ),  $\tau_1, \tau_2, \dots, \tau_n$  are terms,  $\chi$  is a variable, and  $g$  is a variable assignment, then

- (a)  $M, w, g \models \phi \tau_1 \dots \tau_n$  iff  $\langle [\tau_1]^{M,g}, \dots, [\tau_n]^{M,g} \rangle \in V(\phi, w)$ .
- (b)  $M, w, g \models \tau_1 = \tau_2$  iff  $[\tau_1]^{M,g} = [\tau_2]^{M,g}$ .
- (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  $M, w, g \not\models A$  or  $M, w, g \models B$ .
- (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)  $M, w, g \models \forall \chi A$  iff  $M, w, g' \models A$  for all  $\chi$ -variants  $g'$  of  $g$ .
- (i)  $M, w, g \models \exists \chi A$  iff  $M, w, g' \models A$  for some  $\chi$ -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  $wRv$ .

$A$  is **true at  $w$  in  $M$**  iff  $M, w, g \models A$  for every assignment function  $g$  for  $M$ .

Let’s return to the model from above, and let’s add the information that the domain of individuals consists of just Aristotle and Boethius. That is, let  $M$  be the following model:

$$\begin{aligned}
 W &= \{w, v\} \\
 R &= \{\langle w, w \rangle, \langle w, v \rangle\} \\
 D &= \{\text{Aristotle}, \text{Boethius}\} \\
 V(a) &= \text{Aristotle} \\
 V(F, w) &= \{\text{Aristotle}, \text{Boethius}\} \\
 V(F, v) &= \{\text{Boethius}\}
 \end{aligned}$$

This isn't a complete specification of a model because I haven't assigned a meaning to names and predicates other than  $a$  and  $F$ , but we have enough information to determine the truth-value of any  $\mathcal{L}_{MP}$ -sentence whose only non-logical vocabulary are  $a$  and  $F$ .

We can, for example, verify that  $Fa$  is true at  $w$  in  $M$ . A sentence is true at  $w$  in  $M$  iff it is true at  $w$  in  $M$  relative to every assignment function  $g$ . By clause (a) of definition 10.2,  $Fa$  is true at  $w$  in  $M$  relative to  $g$  iff  $[a]^{M,g}$  is a member of  $V(F, w)$ . Since  $a$  is a name,  $[a]^{M,g}$  is  $V(a)$ . And  $V(a)$  is Aristotle. So  $Fa$  is true at  $w$  relative to  $g$  iff Aristotle is a member of  $V(F, w)$ . We know that  $V(F, w)$  is  $\{\text{Aristotle}, \text{Boethius}\}$ . Aristotle evidently is a member of  $\{\text{Aristotle}, \text{Boethius}\}$ . So  $Fa$  is true at  $w$  in  $M$ , relative to any assignment  $g$ .

We can also verify that  $\Box Fa$  is false at  $w$ . By clause (j) of definition 10.2,  $\Box Fa$  is true at  $w$  (in  $M$  relative to  $g$ ) iff  $Fa$  is true (in  $M$  relative to  $g$ ) at all worlds accessible from  $w$ . And  $Fa$  is false at  $v$  because Aristotle is not a member of  $\{\text{Boethius}\}$ .

### Exercise 10.1

Which of the following sentences are true at  $w$  in  $M$ ?

- (a)  $\neg Fa \rightarrow Fa$
- (b)  $\Box \exists x Fx$
- (c)  $\Box \forall x Fx$
- (d)  $\exists x \Box Fx$
- (e)  $\forall x \Box Fx$
- (f)  $\forall x (\Box Fx \rightarrow \Box \Box Fx)$

Validity is truth at all worlds in all models of a certain kind. A sentence is **CK-valid** iff it is true at all worlds in all constant-domain Kripke models. 'C' comes from 'constant domains'; 'K' indicates that we have put no constraints on the accessibility relation. We get stronger concepts of validity – stronger logics – if we require the accessibility relation to be reflexive, or transitive, or euclidean, etc.

It is not hard to see that every sentence that is valid in classical predicate logic is CK-valid. Similarly, every K-valid sentence is CK-valid. We also get some new interaction

principles between modal operators and quantifiers. For example, consider the following schema, known as the *Barcan Formula*, after Ruth Barcan Marcus.

$$(BF) \quad \forall x \Box A \rightarrow \Box \forall x A$$

**Observation 10.1:** All instances of (BF) are CK-valid.

*Proof.* Suppose a sentence  $\forall x \Box A$  is true at some world  $w$  in some constant-domain model  $M$  relative to some assignment  $g$ . By clause (h) of definition 10.2, it follows that  $\Box A$  is true at  $w$  relative to every  $x$ -variant  $g'$  of  $g$ . By clause (j) of definition 10.2, it follows that  $A$  is true at every world  $v$  accessibility from  $w$  relative to every  $x$ -variant  $g'$  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  $g$ .

We've shown that whenever  $\forall x \Box A$  is true at some world  $w$  in some model  $M$  relative some assignment  $g$ , then  $\Box \forall x A$  is also true at  $w$  in  $M$  relative to  $g$ . By clause (f) of definition 10.2, it follows that  $\forall x \Box A \rightarrow \Box \forall x A$  is true at every world in every model relative to every assignment.  $\square$

Instead of working through definition 10.2, we can use trees to test if a sentence is CK-valid. The tree rules for CK are all the rules for K (from chapter 3) together with all the rules for standard predicate logic, with an added world parameter on each node that is held fixed when applying a rule from predicate logic. (In the predicate logic rules, a name counts as 'old' if it already occurs on the relevant branch, no matter at which world.)

To get a complete proof system, we need one further identity rule, reflecting the fact that the reference of a name does not vary from world to world:

Identity Invariance

$$\begin{array}{c} \eta_1 = \eta_2 \quad (\omega) \\ \vdots \\ \eta_1 = \eta_2 \quad (\nu) \\ \uparrow \\ \text{old} \end{array}$$

Here is a tree proof for a simple instance of the Barcan Formula,  $\forall x \Box Fx \rightarrow \Box \forall x Fx$ .

1.	$\neg(\forall x\Box Fx \rightarrow \Box\forall xFx)$	(w) (Ass.)
2.	$\forall x\Box Fx$	(w) (1)
3.	$\neg\Box\forall xFx$	(w) (1)
4.	$wRv$	(3)
5.	$\neg\forall xFx$	(v) (3)
6.	$\neg Fa$	(v) (5)
7.	$\Box Fa$	(w) (2)
8.	$Fa$	(v) (7,4)
	x	

And here is a proof of  $\forall x\forall y(x=y \rightarrow \Box x=y)$ , the “necessity of identity”:

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) (4, 5, LL)
7.	$wRv$	(6)
8.	$b \neq b$	(v) (6)
9.	$b=b$	(v) (SI)
	x	

### Exercise 10.2

Use the tree method to show that the following sentences are CK-valid.

- (a)  $\Box\forall xFx \rightarrow \forall x\Box Fx$
- (b)  $\exists x\Box Fx \rightarrow \Box\exists xFx$
- (c)  $\forall x\Box(Fx \wedge Gx) \rightarrow \Box\forall xFx$
- (d)  $\Box\Diamond\exists xFx \rightarrow \Box\exists x\Diamond(Fx \vee Gx)$
- (e)  $\forall x\Box\exists y y=x$
- (f)  $\forall x\forall y(x \neq y \rightarrow \Box x \neq y)$

**Exercise 10.3**

The following sentences are CK-invalid. Can you describe a countermodel for each? (It may help to construct a tree and inspect its open branches.)

- (a)  $\Diamond \exists x Fx \rightarrow \Diamond \exists x (Fx \wedge Gx)$
- (b)  $\Box \exists x Fx \rightarrow \exists x \Box Fx$
- (c)  $\forall x \forall y ((\Diamond Fx \wedge \Diamond \neg Fy) \rightarrow x \neq y)$
- (d)  $\forall x \Box (Px \rightarrow Qx) \rightarrow \forall x (Px \rightarrow \Box Qx)$

There are also axiomatic calculi for CK. We can, for example, combine the axiom schemas and rules of classical predicate logic with those of K, and add two new schemas: the Barcan Formula (BF) and the “necessity of distinctness”,

$$(ND) \quad \forall x \forall y (x \neq y \rightarrow \Box x \neq y).$$

As I mentioned above, stronger logics can be defined by putting constraints on the accessibility relation. For example, the system **CT** is the set of  $\mathcal{L}_{MP}$ -sentences that are valid in the class of constant-domain models with a reflexive accessibility relation. **CS4** is the set of  $\mathcal{L}_{MP}$ -sentences that are valid in the class of constant-domain models with a reflexive and transitive accessibility relation. And so on.

Properties of the accessibility relation still correspond to modal schemas, just as in chapter 3: (T) corresponds to reflexivity, (4) to transitivity, (G) to convergence, etc. Recall that a schema *corresponds* to a property of the accessibility relation if the schema is valid in all and only the frames in which the accessibility relation has that property. A *frame* is a model without an interpretation function. In the present context, a frame therefore consists of two non-empty sets  $W$  and  $D$  and a relation  $R$  on  $W$ .

We can still use the tree method or the axiomatic method to test for validity in logics stronger than CK. To test for CT-validity, for example, we would add the Reflexivity rule to the tree rules for CK. To test for CS4-validity, we would add the Reflexivity and Transitivity rules. We can get an axiomatic calculus for CT by adding the (T)-schema to the calculus for CK; for CS4, we can add (T) and (4). And so on for other systems.

(Curiously, though, this doesn’t always work. You may remember S4.2 as the set of  $\mathcal{L}_M$ -sentences that are valid in the class of reflexive, transitive, and convergent Kripke models. Reflexivity corresponds to (T), transitivity to (4), and convergence to (G). If we add these schemas to the axiomatic calculus for K, we get a sound and complete calculus for S4.2. But if we add the schemas to the calculus for CK, the resulting calculus is not complete for CS4.2. There are  $\mathcal{L}_{MP}$ -sentences that are valid in the class of reflexive,

transitive, and convergent constant-domain models that can't be derived.)

## 10.2 Quantification and existence

We have assumed that the domain of individuals is the same for every world. This may seem problematic.

Earlier today I was baking bread. Let's call the loaf of bread that I made Loafy. Intuitively, Loafy could have failed to exist. I could have decided not to bake bread. Even if determinism is true, we can consider worlds at which the laws of nature or the origin of the universe are different. In many of these worlds, there are no humans, and no loafs of bread. So we should allow for worlds at which Loafy doesn't exist.

If we use  $b$  as a name for Loafy, we can arguably express Loafy's existence as

$$\exists x x = b.$$

Why might this express that Loafy exists? Consider a scenario in which Loafy does exist. In that scenario, there is some thing  $x$  which is identical to Loafy (namely, Loafy). Conversely, consider a scenario in which Loafy does not exist. In that scenario, there is no thing  $x$  which is identical to Loafy. So  $\exists x x = b$  is true in all and only the scenarios in which Loafy exists.

Now we can sharpen the above worry. Intuitively, it could have been the case that Loafy doesn't exist. So  $\Diamond \neg \exists x x = b$  is true, on a suitable understanding of the diamond. But in constant-domain semantics, that sentence is a contradiction: it is false at every world in every model.

A converse problem arises if we think that something could have existed that doesn't actually exist. For example, let's assume that there could have been unicorns. If we interpret the predicate  $U$  as ' $-$  is a unicorn' and the box as a suitable kind of circumstantial necessity,  $\Box \forall x \neg Ux$  should then be false. But let's also assume that no individual in our world could have been a unicorn. So  $\forall x \Box \neg Ux$  is true. We then have a counterexample to the Barcan Formula  $\forall x \Box A \rightarrow \Box \forall x A$ . And all instances of the Barcan Formula are valid in constant-domain semantics.

### Exercise 10.4

The **Converse Barcan Formula** is the schema  $\Box \forall x A \rightarrow \forall x \Box A$ . All instances of the Converse Barcan Formula are CK-valid. Explain why Loafy's possible non-existence seems to provide a counterexample to the Converse Barcan Formula.

### Exercise 10.5

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)  $\exists x \Diamond A \rightarrow \Diamond \exists x A$
- (4)  $\exists x \Box A \rightarrow \Box \exists x A$

- (a) Are any of (1)–(4) (schematically) equivalent to the Barcan Formula or the Converse Barcan Formula (given the duality of  $\Box$  and  $\Diamond$ , of  $\forall x$  and  $\exists x$ , and the standard truth-tables for propositional connectives)?
- (b) Which of these schemas do you think are intuitively valid on a metaphysical interpretation of the box and the diamond?

An obvious response to these problems is to replace constant-domain semantics with a semantics in which the domain of individuals can vary from world to world. We will explore this option in the following section. First I want to mention two other lines of response.

Some philosophers have argued that we should bite the bullet: we are simply mistaken when we judge that Loafy could have failed to exist, or that anything could have existed that doesn't actually exist. In temporal logic, biting the bullet means to accept that anything that has ever existed still exists today, and that anything that exists today has always existed and is always going to exist. In epistemic logic, biting the bullet means to accept that nobody can be unsure or ignorant about which individuals exists: if something exists, nobody can fail to know that it exists, nor can anyone believe that an individual exists that doesn't really exist.

A different response is to break the link between quantification and existence.  $\exists x$  is traditionally called an "existential" quantifier, and pronounced 'there is an  $x$ ' or 'there exists an  $x$ '. But  $\mathcal{L}_{MP}$  is a made-up language. We can make its symbols mean whatever we want. We can give a different interpretation of  $\exists x$  so that 'Loafy exists' can't be translated as  $\exists x x = b$ .

One alternative to the standard interpretation of quantifiers is associated with the Austrian philosopher Alexius Meinong. Meinong observed that when we describe beliefs, plans, hopes, or fears, we often seem to refer to non-existent objects. We might say that someone is afraid of *a ghost*, or that they are searching for *a golden mountain* – even though there are no ghosts or golden mountains. According to Meinong, people who are

searching for a golden mountain are really searching for *something*. That something is a golden mountain. But it is not an existent golden mountain. Meinong concluded that besides existent mountains, there are also non-existent mountains.

Quantifiers that range over both existent and non-existent individuals are sometimes called *Meinongian*. If the  $\mathcal{L}_{MP}$ -quantifiers are Meinongian, then clearly  $\exists x x = b$  does not translate ‘Loafy exists’.

Meinong’s postulation of non-existent individuals raises difficult questions. Suppose you are searching for a golden mountain. You probably don’t have any firm views about the mountain’s height. You are not looking for a mountain that is exactly 2000 meters tall, nor are you looking for a mountain that is exactly 2100 meters tall. On the Meinongian account, there is a genuine mountain that you are looking for. It is a mountain that is not 2000 meters tall, not 2100 meters tall, and doesn’t have any other particular height either. But how could there be a mountain without any particular height? Besides, it also doesn’t seem right to say that you are looking for a peculiar “mountain” that doesn’t have any height and doesn’t exist. Intuitively, you are looking for an *existent* mountain that *does* have a height.

A more straightforward alternative to the standard interpretation of quantifiers is the *possibilist* interpretation. Here we assume that  $\forall x$  and  $\exists x$  range not only over things that exist at the world at which the quantifiers are interpreted, but over everything that exists at any possible world. On this interpretation, too,  $\exists x x = b$  no longer states that Loafy exists. It merely states that Loafy could have existed, in an unrestricted sense of ‘could’. Constant-domain semantics then only assumes that the set of individuals that exist at some world or other does not vary from world to world.

One downside of the possibilist interpretation is that it goes against the “internalist” spirit of modal logic. As we saw in section 9.2, one of the key features of modal logic is that it looks at the structure of worlds from the inside, from the perspective of a particular world, with only the modal operators providing (incomplete) access to other worlds. Possibilist quantifiers would provide unrestricted access to the inhabitants of other worlds.

Let’s set aside these alternatives and see how constant-domain semantics could be changed to allow for variable domains.

### 10.3 Variable-domain semantics

In variable-domain models, every world  $w$  is associated with its own individual domain  $D_w$ . Loafy the bread may be a member of  $D_w$  but not of  $D_v$ . Quantifiers range over the

individuals in the local domain of the world at which they are interpreted:  $\exists xFx$  is true at  $w$  iff  $Fx$  is true (at  $w$ ) of some individual in  $D_w$ .

Here is our revised definition of an  $\mathcal{L}_{MP}$ -model.

### Definition 10.3

A **variable-domain Kripke model** for  $\mathcal{L}_{MP}$  is a structure  $M$  consisting of

1. a non-empty set  $W$  (the “worlds”),
2. a binary (“accessibility”) relation  $R$  on  $W$ ,
3. for each world  $w$ , a non-empty set  $D_w$  (of “individuals”), 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  $w$  a set of  $n$ -tuples from  $D_w$ .

To complete the semantics, we need to explain how  $\mathcal{L}_{MP}$ -sentences are interpreted relative to any given world in a variable-domain model. This raises a problem.

Since Loafy could have failed to exist, we want to have models in which  $\Diamond \neg \exists x x = b$  is true at some world  $w$ . It follows that  $\neg \exists x x = b$  is true at some world  $v$  accessible from  $w$ . Intuitively,  $v$  is a world at which Loafy doesn’t exist. The problem is that we need to explain how a sentence that contains a name (here,  $b$ ) should be interpreted at a world (here,  $v$ ) where the thing that’s picked out by the name doesn’t exist.

In the case of  $\neg \exists x x = b$ , the sentence should come out true. Other cases are less clear. What about  $b = b$ ? Is Loafy identical to Loafy at  $v$ , where Loafy doesn’t exist? What about  $Fb$ ,  $\neg Fb$ , or  $Fb \vee \neg Fb$ ? Is Loafy delicious at  $v$ ? Is Loafy not delicious at  $v$ ? Is Loafy either delicious or not delicious at  $v$ ?

These questions are discussed not just in modal logic, but also in a branch of non-modal logic called **free logic**. Free logic differs from classical predicate logic by dropping the assumption that every name has a referent. The assumption is, after all, not true for names in natural language.

Consider the story of ‘Vulcan’. In the 19th century, it was observed that Mercury’s path around the Sun conforms to Newton’s laws only if there is another, smaller planet between Mercury and the Sun. With the help of Newton’s laws, astronomers calculated the size and position of that planet, and called it Vulcan. But Vulcan was never discovered. Eventually, Mercury’s path was explained by Einstein’s theory of relativity, without assuming any new planets. The name ‘Vulcan’ turned out to be *empty*: it doesn’t refer to anything.

How should we formalize reasoning with empty names? The orthodox answer is that we shouldn't: the function of a name is to pick out an individual; if there is no individual to be picked out, we shouldn't use a name. Proponents of free logic disagree. They hold that we can perfectly well reason with empty names. We then need to answer the same questions that I posed above: if  $b$  is an empty name, how should we interpret  $b = b$ ,  $Fb$ ,  $\neg Fb$ , and  $Fb \vee \neg Fb$ ?

Within free logic, there are broadly three approaches.

The first is Meinongian. It assumes that apparently empty names are not really empty after all; they merely pick out a non-existent individual. Statements with such names are then interpreted as usual:  $Fb$  may be true or false, depending on whether the (non-existent) individual picked out by  $b$  has the property expressed by  $F$ .

Non-Meinongian versions of free logic usually assume that *atomic* sentences with empty names are never true: if  $b$  is empty, then  $Fb$  can't be true. The idea is that predicates express properties, and if something doesn't exist then it doesn't have any properties. For example, it is not true that Vulcan is a planet – as you can see from the fact that Vulcan would not occur on a list of all planets. Nor is it true that Vulcan orbits the sun, or that Vulcan has any particular mass.

What shall we say about  $\neg Fb$  then, if  $b$  is an empty name? In some versions of free logic, the standard semantic rules for complex sentences are applied: since  $Fb$  is not true,  $\neg Fb$  is true, and so is  $Fb \vee \neg Fb$ . Other versions of free logic assume that if  $b$  doesn't refer then neither  $Fb$  nor  $\neg Fb$  is true. Since a sentence is called false iff its negation is true, this means that  $Fb$  and  $\neg Fb$  are neither true nor false. We get a three-valued semantics that can be spelled out in different ways, with different verdicts on sentences like  $Fb \vee \neg Fb$ .

Each version of free logic can be used to give a semantics for modal predicate logic with variable domains. I am going to use the two-valued non-Meinongian approach, mainly because it is the simplest. We will assume that at worlds where Loafy doesn't exist, every atomic sentence involving a name for Loafy is false:  $b = b$  is false,  $Fb$  is also false, but  $\neg Fb$  and  $Fb \vee \neg Fb$  are true.

**Definition 10.4: Variable-domain Kripke semantics**

If  $M = \langle W, R, D, V \rangle$  is a variable-domain Kripke model,  $w$  is a member of  $W$ ,  $\phi$  is an  $n$ -place predicate (for  $n \geq 0$ ),  $\tau_1, \dots, \tau_n$  are terms,  $\chi$  is a variable, and  $g$  is a variable assignment, then

- |     |  |  |
|-----|--|--|
| (a) | $M, w, g \models \phi \tau_1 \dots \tau_n$ | iff $\langle [\tau_1]^{M,g}, \dots, [\tau_n]^{M,g} \rangle \in V(\phi, w)$ .                 |
| (b) | $M, w, g \models \tau_1 = \tau_2$          | iff $[\tau_1]^{M,g} = [\tau_2]^{M,g}$ and $[\tau_1]^{M,g} \in D_w$ .                         |
| (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 $M, w, g \not\models A$ or $M, w, g \models B$ .   |
| (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) | $M, w, g \models \forall \chi A$           | iff $M, w, g' \models A$ for all $\chi$ -variants $g'$ of $g$ for which $g'(\chi) \in D_w$ . |
| (i) | $M, w, g \models \exists \chi A$           | iff $M, w, g' \models A$ for some $\chi$ -variant $g'$ of $g$ for which $g'(\chi) \in D_w$ . |
| (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 $wRv$ .                                 |

$A$  is **true at  $w$  in  $M$**  iff  $M, w, g \models A$  for all assignments  $g$  for  $M$ .

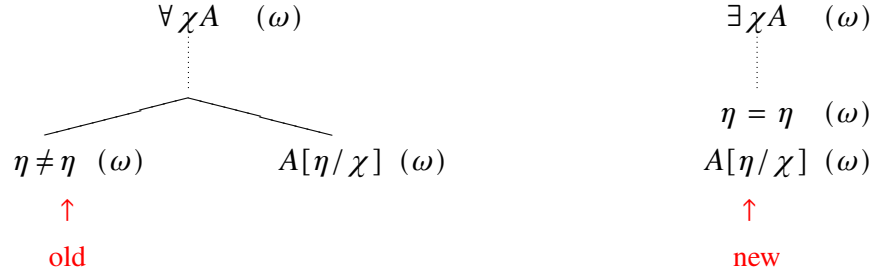
A sentence is **VK-valid** ('V' for 'variable-domain') iff it is true at all worlds in all variable-domain models.

The system VK is weaker than classical predicate logic. Not everything that is valid in classical predicate logic is CK-valid. For example, both  $b = b$  and  $\exists x x = b$  are valid in classical predicate logic, but they are not true at every world in every variable-domain model. If  $V(b)$  is not a member of  $D_w$ , then  $b = b$  and  $\exists x x = b$  are false at  $w$ .

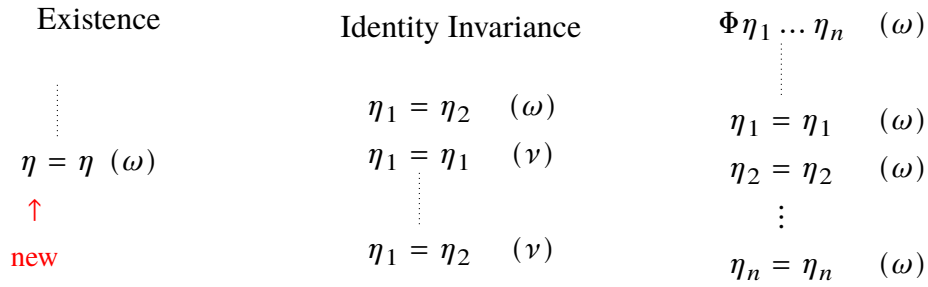
On the other hand, you can check that  $\forall x x = x$  is VK-valid. So we don't just have to revise the rules for identity. We also need to revise the rule of "universal instantiation": from the fact that a universal generalisation like  $\forall x x = x$  is true (at a world, or at all worlds), we can't infer that all its instances are true:  $b = b$  may be false. For another example, consider a world  $w$  where everything is made of chocolate. Let  $F$  express the property of being made of chocolate.  $\forall x Fx$  is true at  $w$ . But we can't infer that Loafy the bread is made of chocolate ( $Fb$ ) at  $w$ , for Loafy may not exist at  $w$ .

In the type of free logic we have adopted, the rule of universal instantiation requires another premise: from  $\forall x A$  we can infer  $A[b/x]$  only if we also know that  $b$  exists – which can be expressed as  $\exists x x = b$ , or even simpler as  $b = b$ , given our assumption that atomic sentences with empty names are always false.

Here are the revised tree rules for VK. I only give the quantifier rules for  $\forall \chi A$  and  $\exists \chi A$ . You can find the rules for  $\neg \forall \chi A$  and  $\neg \exists \chi A$  by converting these into  $\exists \chi \neg A$  and  $\forall \chi \neg A$ , respectively.



We keep the rule for Leibniz's Law. But we replace the Self-Identity and Identity Invariance rules by the following three rules.



The Existence rule reflects our assumption that the domain of individuals is never empty. The unnamed last rule is a rule for expanding atomic nodes. From the assumption that  $Fb$  is true at a world, for example, the rule allows us to infer that  $b$  exists at that world, which can be expressed as  $b = b$ . We then don't need a separate rule of Self-Identity.

### Exercise 10.6

Use the tree method to show that the following sentences are VK-valid.

- (a)  $\exists x \Box Fx \rightarrow \Box \exists x Fx$
- (b)  $\Box \forall x (Fx \rightarrow Gx) \rightarrow (\Box \forall x Fx \rightarrow \Box \forall x Gx)$
- (c)  $\Box \exists x x = x$
- (d)  $\Diamond Fa \rightarrow \Diamond \exists x Fx$
- (e)  $a = b \rightarrow \Box (a = a \rightarrow a = b)$

It is easy to check that the Barcan Formula,  $\forall x \Box A \rightarrow \Box \forall x A$ , and its converse,  $\Box \forall x A \rightarrow \forall x \Box A$ , are invalid in variable-domain semantics. (By this I mean that not all their instances are valid.) In fact, we can now prove that the Barcan formula corresponds to the assumption that whatever exists at an accessible world also exists at the original world, while its converse corresponds to the assumption that whatever exists at a world also exists at all accessible worlds.

**Observation 10.2:**

- (i) (CBF) is valid on a variable-domain frame iff the frame has *increasing domains*, meaning that whenever  $wRv$ , then  $D_w \subseteq D_v$ .
- (ii) (BF) is valid on a variable-domain frame iff the frame has *decreasing domains*, meaning that whenever  $wRv$  then  $D_v \subseteq D_w$ .

*Proof of (i).* Suppose some variable-domain frame  $F$  does not have increasing domains. Then  $F$  has a world  $w$  whose domain  $D_w$  contains an individual  $d$  that does not exist at some  $w$ -accessible world  $v$ . Let  $V$  be an interpretation function on  $F$  so that  $V(F, w) = D_w$  and  $V(F, v) = D_v$ . In the model composed of  $F$  and  $V$ ,  $\Box \forall x Fx$  is true at  $w$ , but  $\forall x \Box Fx$  is false, since  $d$  is not in  $V(F, v)$ . So (CBF) is not true at all worlds in all models based on  $F$ .

In the other direction, suppose (CBF) is not valid on a frame  $F$ . This means that there is a world  $w$  in some model  $M$  based on  $F$  at which some instance of  $\Box \forall x A$  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 of some individual  $d$  in  $D_w$ . But since  $\Box \forall x A$  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 proof of (ii) is similar. □

**Exercise 10.7**

Definition 10.3 requires that every name in every model picks out a possible individual. In that sense, the definition does not allow for genuinely empty names. How could we change definitions 10.3 and 10.4 if we wanted to allow for names that don't pick out anything?

## 10.4 Trans-world identity

In section 9.5 I mentioned an apparent problem with Leibniz' Law. The Law allows us to reason from  $\Box Fa$  and  $a = b$  to  $\Box Fb$ . On some interpretations of the box, this inference looks problematic. In the Superman stories, Lois Lane knows that Superman can fly, and Superman is identical to Clark Kent. Can we infer that Lois knows that Clark Kent can fly?

If we can, we would have to conclude that Lois Lane has inconsistent beliefs, since she also believes that Clark Kent *cannot* fly. She would believe that Clark Kent can't fly,

but also that he can fly. Intuitively, however, Lois's beliefs are perfectly consistent. What she lacks is information, not logical acumen. Her belief worlds are not worlds at which someone can both fly and not fly. Rather, they are worlds at which one person plays the Superman role and a different person plays the Clark Kent role.

Consider also the case of Julius. When we introduce the name 'Julius' for whoever invented the zip, we can be sure that Julius invented the zip. But it would be absurd to think that we have found out who invented the zip merely by making a linguistic stipulation. If before introducing the name 'Julius', we were unsure whether the zip was invented by Benjamin Franklin or Whitcomb L. Judson, the introduction of the new name does nothing to remove our ignorance. There are still epistemically accessible worlds at which the zip was invented by Franklin and others at which it was invented by Judson. Knowing that Julius invented the zip is not the same thing as knowing that Judson invented the zip, even if in fact Julius = Judson.

Similar problems have been argued to arise in the logic of metaphysical modality. Imagine a clay statue, standing on a shelf. Let's call it Goliath. Since Goliath is made of clay, there is also a piece of clay on the shelf, at the exact same spot as the statue. Let's call that piece of clay Lump1. How is Lump1 related to Goliath? We might want to say that they are one and the same thing: Lump1 = Goliath. After all, there is only *one* statue-shaped object on the shelf, not two. But we might also want to say that Lump1 could have had the shape of a bowl, while Goliath could not: if the clay had been formed into a bowl rather than a statue, then Lump1 would have been a bowl, but Goliath, the statue, would not have existed. Goliath is necessarily not a bowl, but Lump1 is not necessarily not a bowl. We have  $\Box \neg Bg$  but not  $\Box \neg Bl$ , even though  $l = g$ .

### Exercise 10.8

Explain why the three examples I just presented also cast doubt on the "necessity of identity",  $\forall x \forall y (x = y \rightarrow \Box x = y)$ .

So perhaps we should give up Leibniz' Law. Semantically, the Law corresponds to the assumption that names are **directly referential**, meaning that the only contribution a name makes to the truth-value of a sentence is its referent. If names are directly referential, and two names have the same referent, then it makes no difference which of them we use: replacing one by the other never affects the truth-value of a sentence.

So far, we have assumed direct reference in both constant-domain and variable-domain semantics. On either account, names are interpreted as simply picking out an individual. It is a matter of debate whether names in ordinary language are directly referential. Some

hold that Lois Lane really has inconsistent beliefs. Others hold that Lois neither believes that Superman can fly nor that Clark Kent can't fly, because the objects of belief or knowledge are never adequately represented by statements involving ordinary names. (This also gets around the Julius problem.) With respect to Lump and Goliath, some simply deny that Lump is identical to Goliath.

We will not descend into these debates. Instead, let's explore how we could change our semantics for  $\mathcal{L}_{MP}$  to block the relevant applications of Leibniz' Law. There are several ways to achieve this. We will only look at one.

The approach we will explore drops the assumption that names are rigid. A name is **rigid** if it picks out the same individual relative to any possible world. Earlier, we assumed that no matter at which world the sentence  $Fa$  is interpreted, the name  $a$  always picks out the same individual,  $V(a)$ . A name like 'Julius', however, seems to be non-rigid. It picks out different individuals relative to different (epistemically) possible worlds. Relative to a world where Benjamin Franklin invented the zip, 'Julius' picks out Benjamin Franklin. Relative to a world where Whitcomb L. Judson invented the zip, the name picks out Whitcomb L. Judson.

Let's assume, then, that a model's interpretation function assigns an individual to each name *relative to each world*. This is equivalent to assuming that each name is interpreted as expressing a function from worlds to individuals, telling us which individual the name picks out relative to any given world. Functions from worlds to individuals are known as **individual concepts**, which is why the present approach is often called **individual concept semantics**.

To motivate this label, return to Lois Lane. When Lois is thinking about Superman, she is thinking about the audacious hero whose superhuman powers she has witnessed on several occasions. When she is thinking about Clark Kent, she is thinking about her shy and awkward colleague. Lois has distinct "concepts" for Superman and Clark Kent, one associated with the Superman role, the other with the Clark Kent role. The two concepts actually pick out the same person because one and the same person plays both the Superman role and the Clark Kent role. We can model each of these roles as a function from worlds to individuals. The Superman role is represented by a function that maps every world to whoever plays the Superman role at that world. The Clark Kent role is represented by a function that maps every world to whoever plays the Clark Kent role at that world. For the world of the Superman stories, both functions return the same individual. For Lois Lane's belief worlds, they return different individuals.

**Exercise 10.9**

What individual concepts might be associated with the names ‘Lumpl’ and ‘Goliath’?

We can easily convert our earlier constant-domain and variable-domain semantics into an individual concept semantics. We first need to change the definition of a model, so that  $V$  assigns individual concepts to names. In variable-domain semantics, we might stipulate that an individual concept never maps a world to an individual that doesn’t exist at the world. We might also want to allow for “partial concepts”: individual concepts that don’t return any value for certain worlds.

It is advisable to give a parallel treatment for names and variables. So we’ll also assume that an assignment function  $g$  interprets each variable as expressing an individual concept. In the truth definition, we replace  $[\tau]^{M,g}$ , by  $[\tau]^{M,w,g}$ , which is defined as the referent of  $\tau$  in  $M$  at  $w$ , relative to  $g$ . (That is, if  $\tau$  is a name, then  $[\tau]^{M,w,g} = V(\tau)(w)$ ; if  $\tau$  is a variable, then  $[\tau]^{M,w,g} = g(\tau)(w)$ .) 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 individual concept it assigns to  $x$ .

The resulting logic of individual concepts has some unexpected features. For example, all instances of the following schema become valid:

$$\Box \exists x A \rightarrow \exists x \Box A$$

To see why, consider the instance  $\Box \exists x Fx \rightarrow \exists x \Box Fx$ . Suppose the antecedent is true at some world in some model. This means that at every accessible world  $v$ , there is at least one individual that is  $F$ . In this case, there are functions that map every accessible world to some individual that is  $F$ . Let  $g'(x)$  be some such function. Relative to  $g'$ ,  $\Box Fx$  is true at  $w$ . So  $\exists x \Box Fx$  is true at  $w$ .

This is widely regarded as problematic. It would suggest that the two readings of ‘something necessarily exists’ are actually equivalent: it is necessary that something or other exists just in case there is something that necessarily exists.

Another problematic feature of individual concept semantics is that the resulting logic has no sound and complete proof procedure. There are no tree rules, or natural deduction rules, or axioms and inference rules that would allow proving all and only the sentences that are true at all worlds in all models of individual concept semantics (no matter if we assume constant or variable domains). It’s not just that no-one has yet found a suitable proof method. One can prove that no such method exists.

Both of these problems can be avoided by putting further constraints on models. We

have assumed that any function from worlds to individuals is a candidate interpretation for a name or a variable. Relative to an assignment function, a variable may pick out Donald Trump in one world, the Eiffel tower in another, a fried egg in a third, and so on. Ordinary concepts are not that gerrymandered. We might therefore identify a certain subset of all individual concepts as “eligible” for being expressed by names or variables. If this is done sensibly,  $\Box\exists xA \rightarrow \exists x\Box A$  becomes invalid, and complete proof methods become available.

#### Exercise 10.10

The following line of thought may be attributed to Descartes. “I am certain that I exist, but not that my body exists. [After all, it could turn out that I am a disembodied soul.] Therefore: I am not my body.” Translate the argument into  $\mathcal{L}_{MP}$ . Is it CK-valid? Is it VK-valid? Do you find it convincing?

#### Exercise 10.11

The following sentence sounds contradictory.

Some ticket will win, but I don’t know if it will win.

Translate the sentence into  $\mathcal{L}_{MP}$ . Explain why its apparent contradictoriness poses a problem for accounts on which variables are treated as directly referential.

#### Exercise 10.12

In individual concept semantics, both the necessity of identity and the necessity of distinctness are invalid. How could we change the semantics to make the necessity of identity valid, but not the necessity of distinctness? (Assume constant domains.)



# 11 Answers to the Exercises

## Chapter 1

### Exercise 1.1

(a), (c), and (d) are  $\mathcal{L}_M$ -sentences, (b), (e), and (f) are not.

### Exercise 1.2

Here is a combined truth table for all the classical connectives:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \rightarrow B$	$A \leftrightarrow B$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

### Exercise 1.3

An operator  $O$  is truth-functional if you can figure out the truth-value of  $Op$  from the truth-value of  $p$ .

(c) and (g) are truth-functional; (a), (b), (d), and (e) are not truth-functional.

(f) is truth-functional if God is omniscient (and infallible); it is also truth-functional if God doesn't exist, or if God believes all and only false things; otherwise (f) is not truth-functional.

### Exercise 1.4

- (a)  $\Diamond p$   $p$ : I offended the principal.
- (b)  $\neg \Diamond p$   $p$ : It is raining.
- (c)  $\Diamond p$   $p$ : There is life on Mars.
- (d)  $\Box(p \rightarrow q)$   $p$ : The murderer escaped through the window;  $q$ : There are traces on the ground.

- (e)  $\Diamond(p \wedge q)$   $p$ : The murderer escaped through the window;  $q$ : There are traces on the ground.

#### Exercise 1.5

- (a)  $\Box p$   $p$ : I go home.  
 (b)  $\neg\Box p$   $p$ : You come.  
 (c)  $\neg\Diamond p$   $p$ : You have another beer.  
 (d)  $\Box(\neg p \rightarrow q)$   $p$ : You have a ticket;  $q$ : You pay a fine.

#### Exercise 1.6

- (a)  $\Diamond p$   $p$ : I study architecture.  
 (b)  $\Diamond p$   $p$ : The bridge collapses.  
 (c)  $\neg\Diamond(p \wedge q)$   $p$ : You are talking to me from the kitchen;  $q$ : I hear you.  
 (d)  $p \rightarrow \Diamond q$   $p$ : You have a smartphone;  $q$ : You use an electronic ticket.

#### Exercise 1.7

The proposed definition is equivalent to definition 1.2 for many languages, but not for all. Consider the sentence  $\exists x\exists y\neg(x = y)$  in the language of predicate logic. If we treat the identity symbol as logical, this sentence contains no non-logical expressions 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's not 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.

#### Exercise 1.8

The following pairs are duals: (a) and (c), (b) and (d), (e) and (g), (f) and (h), (i) and (k), (l) and (l), (m) and (m).

#### Exercise 1.9

(b) and (e) are equivalent to  $\Diamond\Diamond\neg p$ , (a), (c), and (d) are not.

As a rule, you can always replace a modal operator by its dual, insert a negation on both sides, and remove any double negations to get an equivalent sentence.

#### Exercise 1.10

(b) and (d)

**Exercise 1.11**

(a)  $\Diamond\Diamond A \rightarrow \Diamond A$ , (b)  $\Diamond\Box A \rightarrow \Box A$ , (c)  $\Box A \rightarrow \Diamond A$ .

**Exercise 1.12**

(a)  $\neg\Box p \wedge \neg\Box\neg p$ ; (b)  $\Diamond p \wedge \Diamond\neg p$ ; (c)  $\neg\forall p \wedge p$ . The last answer assumes that every necessary proposition is true. Without that assumption there is no answer to (c).

**Exercise 1.13**

- (a) All of them.
- (b) Only (K) and (CPL).
- (c) All except (T).
- (d) All of them.

**Exercise 1.14**

(a)

1.  $\Box p \rightarrow p$  (T)
2.  $\Box(\Box p \rightarrow p)$  (1, Nec)

(b)

1.  $p \rightarrow (q \rightarrow (p \wedge q))$  (CPL)
2.  $\Box(p \rightarrow (q \rightarrow (p \wedge q)))$  (1, Nec)
3.  $\Box(p \rightarrow (q \rightarrow (p \wedge q))) \rightarrow (\Box p \rightarrow \Box(q \rightarrow (p \wedge q)))$  (K)
4.  $\Box p \rightarrow \Box(q \rightarrow (p \wedge q))$  (2, 3, CPL)
5.  $\Box(q \rightarrow (p \wedge q)) \rightarrow (\Box q \rightarrow \Box(p \wedge q))$  (K)
6.  $\Box p \rightarrow (\Box q \rightarrow \Box(p \wedge q))$  (4, 5, CPL)
7.  $(\Box q \wedge \Box p) \rightarrow \Box(p \wedge q)$  (6, CPL)

(c)

1.  $\neg\Diamond\neg p \leftrightarrow \Box\neg\neg p$  (Dual)
2.  $\neg\neg\Diamond\neg p \leftrightarrow \neg\Box\neg\neg p$  (1, CPL)
3.  $\Diamond\neg p \leftrightarrow \neg\Box\neg\neg p$  (2, CPL)
4.  $\neg\neg p \rightarrow p$  (CPL)
5.  $\Box(\neg\neg p \rightarrow p)$  (4, Nec)
6.  $\Box(\neg\neg p \rightarrow p) \rightarrow (\Box\neg\neg p \rightarrow \Box p)$  (K)
7.  $\Box\neg\neg p \rightarrow \Box p$  (5, 6, CPL)
8.  $p \rightarrow \neg\neg p$  (CPL)
9.  $\Box(p \rightarrow \neg\neg p)$  (8, Nec)
10.  $\Box(p \rightarrow \neg\neg p) \rightarrow (\Box p \rightarrow \Box\neg\neg p)$  (K)
11.  $\Box p \rightarrow \Box\neg\neg p$  (9, 10, CPL)
12.  $\Box\neg\neg p \leftrightarrow \Box p$  (7, 11, CPL)
13.  $\neg\Box\neg\neg p \leftrightarrow \neg\Box p$  (12, CPL)
14.  $\Diamond\neg p \leftrightarrow \neg\Box p$  (3, 13, CPL)

### Exercise 1.15

In an axiomatic calculus, every line in a proof is either an axiom or follows from an earlier line by one of the rules. (Nec) therefore assumes that whenever a sentence  $A$  is *provable in the axiomatic calculus*, then it is necessarily true (reading the box as ‘it is necessary that’).

The rules of the axiomatic calculus cannot be used to directly derive assumptions from arbitrary premises. To show that  $A$  entails  $B$ , you have to prove  $A \rightarrow B$ .

## Chapter 2

### Exercise 2.1

Consider a scenario in which (say) it is raining at some worlds and not raining at others. Let  $p$  express that it is raining. In this scenario, under this interpretation,  $\Diamond p$  is true, because  $p$  is true at some world. But  $\Box p$  is false, because  $p$  is not true at all worlds. So there are conceivable scenarios and interpretations that render  $\Diamond p$  true and  $\Box p$  false.

### Exercise 2.2

(b), (e), and (f) are true at  $w_1$ , the others false.

### Exercise 2.3

$\Diamond p \rightarrow (q \vee \Diamond \Box p)$  is true at both worlds.

### Exercise 2.4

The two definitions are not equivalent, as can be seen from the fact that the definition proposed in the exercise would render  $p \models \Box p$  true. Whenever  $p$  is true at every world in a model then (by definition 2.2)  $\Box p$  is also true at every world in the model. Definition 2.4 renders  $p \models \Box p$  false, since there are models in which  $p$  is true at some worlds and not at others.

### Exercise 2.5

By definition 2.3, a sentence is valid iff it is true at every world in every model. Suppose for reductio that  $\Box p \rightarrow \Diamond p$  is false at some world  $w$  in some model. By definition 2.2,  $\Box p$  is then true at  $w$  and  $\Diamond p$  false. But if  $\Diamond p$  is false at  $w$  then (by definition 2.2)  $p$  is false at every world in the model. And then  $\Box p$  isn't true at  $w$  (by definition 2.2). Contradiction.

### Exercise 2.6

Suppose  $A$  is valid – true at all worlds in all models (definition 2.3). It follows that in any given model,  $A$  is true at every world. By definition 2.2, it follows that  $\Box A$  is true at every world in any model.

### Exercise 2.7

$p \rightarrow \Box p$  is false at world  $w$  in the model(s) given by  $W = \{w, v\}$ ,  $V(p) = \{w\}$ .

This shows that the *truth* of  $p$  (at a world in a model) does not entail the truth of  $\Box p$  (at the world in the model), even though the *validity* of  $p$  entails the validity of  $\Box p$ , as per the previous exercise.

### Exercise 2.8

Assume  $\models A \rightarrow B$ . Then there is no world in any model at which  $A$  is true and  $B$  is false. So if  $A$  is true at every world in a model, then  $B$  is also true at every world in the model. It follows that  $\Box A \rightarrow \Box B$  is true at every world in every model.

### Exercise 2.9

(a) Target:  $p \rightarrow q$

1.  $\neg(p \rightarrow q)$  (w) (Ass.)
2.  $p$  (w) (1)
3.  $\neg q$  (w) (1)

Countermodel:  $W = \{w\}$ ,  $V(p) = \{w\}$ ,  $V(q) = \emptyset$ .

(b) Target:  $p \rightarrow \Box(p \vee q)$

1.  $\neg(p \rightarrow \Box(p \vee q))$  (w) (Ass.)
2.  $p$  (w) (1)
3.  $\neg\Box(p \vee q)$  (w) (1)
4.  $\neg(p \vee q)$  (v) (3)
5.  $\neg p$  (v) (4)
5.  $\neg q$  (v) (4)

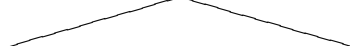
Countermodel:  $W = \{w, v\}$ ,  $V(p) = \{w\}$ ,  $V(q) = \emptyset$ .

(c) Target:  $\Box p \vee \Box \neg p$

1.  $\neg(\Box p \vee \Box \neg p)$  (w) (Ass.)
2.  $\neg\Box p$  (w) (1)
3.  $\neg\Box \neg p$  (w) (1)
4.  $\neg p$  (v) (2)
5.  $\neg\neg p$  (u) (3)
6.  $p$  (u) (5)

Countermodel:  $W = \{w, v, u\}, V(p) = \{u\}$ .

(d) Target:  $\Diamond(p \rightarrow q) \rightarrow (\Diamond p \rightarrow \Diamond q)$

1.	$\neg(\Diamond(p \rightarrow q) \rightarrow (\Diamond p \rightarrow \Diamond q))$	(w)	(Ass.)
2.	$\Diamond(p \rightarrow q)$	(w)	(1)
3.	$\neg(\Diamond p \rightarrow \Diamond q)$	(w)	(1)
4.	$\Diamond p$	(w)	(3)
5.	$\neg\Diamond q$	(w)	(3)
6.	$p \rightarrow q$	(v)	(2)
7.	$p$	(u)	(4)
8.	$\neg q$	(w)	(5)
9.	$\neg q$	(v)	(5)
10.	$\neg q$	(u)	(5)
			
11.	$\neg p$	(v)	(6)
12.	$q$	(v)	(6)
		x	

Countermodel:  $W = \{w, v, u\}, V(p) = \{u\}, V(q) = \emptyset$ .

(e)  $\Box\Diamond p \rightarrow p$

1.	$\neg(\Box\Diamond p \rightarrow p)$	(w)	(Ass.)
2.	$\Box\Diamond p$	(w)	(1)
3.	$\neg p$	(w)	(1)
4.	$\Diamond p$	(w)	(2)
5.	$p$	(v)	(4)
6.	$\Diamond p$	(v)	(2)
7.	$p$	(u)	(6)
8.	$\Diamond p$	(u)	(2)
9.	$p$	(t)	(8)

The tree grows forever. The target sentence isn't valid, but the tree method only gives us an infinite countermodel. In such a case, it may be useful to read off a model from an incomplete version of the tree and manually check whether it is a genuine countermodel. The model determined by the first five nodes of the present

tree is  $W = \{w, v\}$ ,  $V(p) = \{v\}$ , and you can confirm that it is a countermodel to the target sentence.

If you read off a model from an *incomplete* tree, you can't be sure that it is a countermodel for the target sentence. You must always double-check!

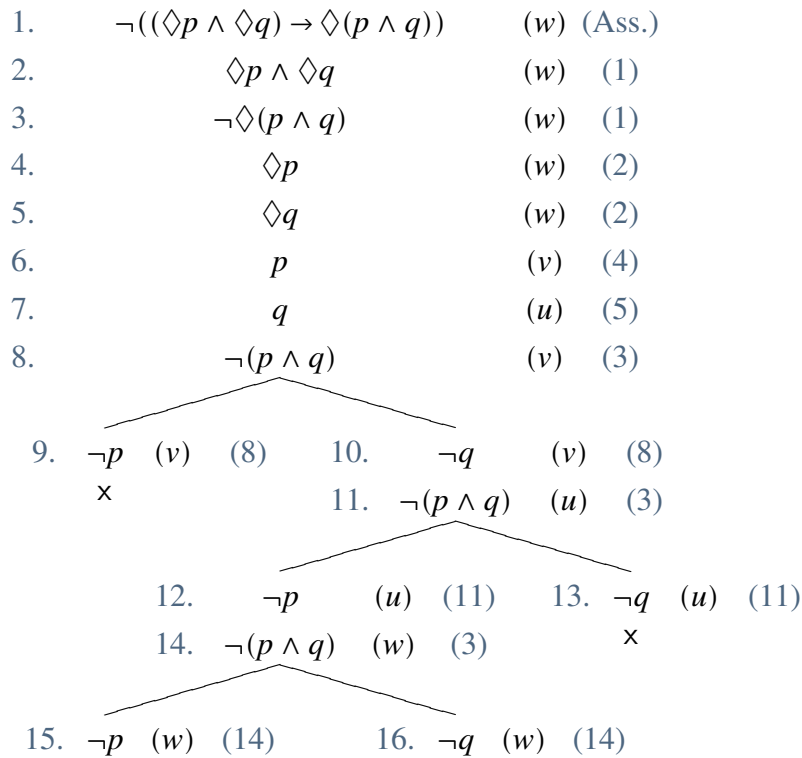
### Exercise 2.10

You can enter the schemas at [umsu.de/trees](https://umsu.de/trees). After entering a formula, tick the checkbox for 'universal (S5)'. Alternatively, follow these links: (K), (T), (4), (5),

### Exercise 2.11

(a), (b), (c) and (e) are valid. You can find the trees at [umsu.de/trees](https://umsu.de/trees) (Remember to tick the checkbox for 'universal (S5)') or by following these links: (a), (b), (c), (e).

(d) and (f) are invalid. Here is a tree for (d):



We can choose either of the open branches to read off a countermodel. In fact, here we get the same countermodel no matter which open branch we choose:  $W = \{w, v, u\}$ ,  $V(p) = \{v\}$ ,  $V(q) = \{u\}$ .

A tree for (e) might begin like this:

1.  $\neg(\Box\Diamond p \rightarrow \Diamond\Box p)$  (w) (Ass.)
2.  $\Box\Diamond p$  (w) (1)
3.  $\neg\Diamond\Box p$  (w) (1)
4.  $\Diamond p$  (w) (2)
5.  $p$  (v) (4)
6.  $\neg\Box p$  (w) (3)
7.  $\neg p$  (u) (6)
8.  $\Diamond p$  (v) (2)
9.  $p$  (s) (8)
10.  $\neg\Box p$  (v) (3)
11.  $\neg p$  (t) (10)

The tree grows forever. The model determined by the first seven nodes of the present tree is  $W = \{w, v, u\}$ ,  $V(p) = \{v\}$ . It is a countermodel to the target sentence.

### Exercise 2.12

By observation 1.1,  $A_1, \dots, A_n$  entail  $B$  iff  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$  is valid. To show that  $A_1, \dots, A_n$  entail  $B$  we could therefore draw a tree for  $(A_1 \wedge \dots \wedge A_n) \rightarrow B$ . In practice, we can save a few steps by starting the tree with multiple assumptions: one for each of the premises  $A_1, \dots, A_n$ , and one for the negated conclusion  $\neg B$ . (All of these are assumed to be true at world  $w$ .) If the tree closes,  $A_1, \dots, A_n$  entail  $B$ .

To show that  $A$  and  $B$  are equivalent, we can draw a tree for  $A \leftrightarrow B$ .

## Chapter 3

### Exercise 3.1

- (a) One answer:  $A$  is physically necessary iff  $A$  is true at all worlds that are compatible with the laws of physics. A possibly better answer:  $A$  is physically necessary iff  $A$  is true at all worlds that are compatible with the laws of physics and the current state of the universe.
- (b) A simple, if somewhat uninformative answer: We know that  $A$  iff  $A$  is true at all worlds that are compatible with our knowledge.
- (c) It is true that  $A$  iff  $A$  is true at all worlds that are identical to the actual world.

### Exercise 3.2

$v$  has access to no world. So any sentence  $A$  is true at *all* (zero) worlds accessible from  $v$ .

If this seems strange, remember that  $\Box A$  is equivalent to  $\neg \Diamond \neg A$ . And  $\Diamond \neg A$  means that there's an accessible world where  $\neg A$  is true. If there are no accessible worlds, then this is false. So  $\neg \Diamond \neg A$  is true.

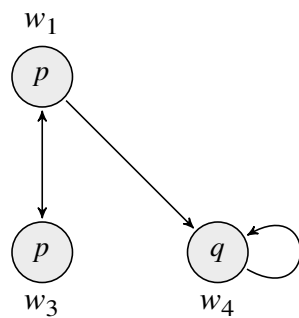
### Exercise 3.3

- (a)  $w_1, w_2$ , and  $w_3$ ; (b)  $w_3$ ; (c)  $\neg$ ; (d)  $w_1, w_2$  and  $w_4$ ; (e) all.

### Exercise 3.4

There are infinitely many correct answers for each world. For example:  $w_1 : \Diamond \Box p$ ,  $w_2 : \neg p \wedge \neg q$ ,  $w_3 : \Box p$ ,  $w_4 : \Box q$ .

### Exercise 3.5



### Exercise 3.6

- (a) For example:  $W = \{w, v\}$ ,  $R = \{(w, v), (v, w)\}$ ,  $V(p) = \{v\}$ .  $\Box p \rightarrow \Box\Box p$  is false at  $w$ . ( $R = \{(w, v), (v, w)\}$  means that  $R$  relates  $w$  to  $v$  and  $v$  to  $w$  and nothing else to anything else.)
- (b) For example:  $W = \{w, v\}$ ,  $R = \{(w, w), (w, v)\}$ ,  $V(p) = \{w\}$ .  $\Diamond p \rightarrow \Box\Diamond p$  is false at  $w$ .

### Exercise 3.7

For example:  $\Box(p \vee \neg p) \rightarrow (p \vee \neg p)$ .

### Exercise 3.8

By clause (g) of definition 3.2,  $\Box(p \vee \neg p)$  is false at a world  $w$  in a Kripke model only if  $p \vee \neg p$  is false at some world accessible from  $w$ . By clause (d) of definition 3.2,  $p \vee \neg p$  is false at a world only if both  $p$  and  $\neg p$  are false at the world, which by clause (a) means that  $p$  is both true and false at the world. This is impossible. So  $\Box(p \vee \neg p)$  is not false at any world in any Kripke model.

### Exercise 3.9

By definition 3.2,  $\Box p \rightarrow \Diamond p$  is false at a world  $w$  in a Kripke model only if  $\Box p$  is true at  $w$  and  $\Diamond p$  is false at  $w$ . But if  $w$  has access to itself then the truth of  $\Box p$  at  $w$  implies that  $p$  is true at  $w$ , and then  $\Diamond p$  is true at  $w$ . So  $\Box p \rightarrow \Diamond p$  can't be false at any world in any Kripke model in which each world has access to itself.

### Exercise 3.10

Reflexive no, serial no, transitive no, euclidean no, symmetric yes, universal no.

### Exercise 3.11

- (a) Suppose  $R$  is symmetric and transitive, and that  $xRy$  and  $xRz$ . By symmetry,  $yRx$ . By transitivity,  $yRz$ .
- (b) Suppose  $R$  is symmetric and euclidean, and that  $xRy$  and  $yRz$ . By symmetry,  $yRx$ . By euclidicity,  $xRz$ .
- (c) Suppose  $R$  is reflexive and euclidean, and that  $xRy$ . By reflexivity,  $xRx$ . By euclidicity,  $yRx$ .

**Exercise 3.12**

It's true that if  $R$  is symmetric and transitive then  $wRv$  implies  $vRw$  which implies  $wRw$ . But this only shows that every world  $w$  *that can see some world*  $v$  can see itself. Symmetry, transitivity, *and seriality* together imply reflexivity. Symmetry and transitivity alone do not.

**Exercise 3.13**

- (a) Every world has access only to itself.
- (b) No world has access to any world.

**Exercise 3.14**

You can enter the sentences at [umsu.de/trees](https://umsu.de/trees). To check for K-validity, leave all the checkboxes (for 'universal' etc.) empty.

**Exercise 3.15**

You can enter the sentences at [umsu.de/trees](https://umsu.de/trees). To test for K4-validity, check the 'transitive' box. To test for D-validity, check 'serial'. To test for B-validity, check 'symmetric'. To test for T-validity, check 'reflexive'.

## Chapter 4

### Exercise 4.1

Methods A and B are genuine proof methods. Method C is not because there is no simple mechanical check of whether a sentence occurs in some logic textbook.

### Exercise 4.2

Method A is complete, but not sound. Everything that's K-valid is provable with the method, but so is everything that's not K-valid.

Method B is sound, but not complete. Since every instance of  $\Box(A \vee \neg A)$  is K-valid, everything that is provable with method B is K-valid. But many K-valid sentences (e.g.,  $p \rightarrow p$ ) aren't provable with method B.

Method C is neither sound nor complete. It is not sound because many K-invalid sentences figure in logic textbooks. It is not complete because there are infinitely many K-valid sentences almost all of which don't occur in any textbooks.

### Exercise 4.3

For  $A \rightarrow B$ : Suppose  $\beta$  contains a node of the form  $A \rightarrow B (\omega)$  and the branch is split into two, with  $\neg A (\omega)$  appended to one end and  $B (\omega)$  to the other. Since the expanded node is a correct statement about  $M$  under  $f$ , we have  $M, f(\omega) \models A \rightarrow B$ . By clause (e) of definition 3.2, it follows that either  $M, f(\omega) \not\models A$  or  $M, f(\omega) \models B$ . By clause (b), this means that either  $M, f(\omega) \models \neg A$  or  $M, f(\omega) \models B$ . So at least one of the resulting branches also correctly describes  $M$ .

For  $\neg \Diamond A$ : Suppose  $\beta$  contains nodes of the form  $\neg \Diamond A (\omega)$  and  $\omega R v$ , and the branch is extended by adding  $\neg A (v)$ . Since  $\neg \Diamond A (\omega)$  and  $\omega R v$  are correct statement about  $M$  under  $f$ , we have  $M, f(\omega) \models \neg \Diamond A$  and  $f(\omega) R f(v)$ . By clause (b) of definition 3.2,  $M, f(\omega) \models \neg \Diamond A$  implies  $M, f(\omega) \not\models \Diamond A$ . By clause (h), it follows that  $M, f(v) \models \neg A$ . So the extended branch correctly describes  $M$ .

### Exercise 4.4

Yes. The function  $f$  can map both 'w' and 'v' to  $w$ .

### Exercise 4.5

A sentence is K4-valid iff it is true at all worlds in all transitive Kripke models. We only

need to check that the Transitivity rule is sound, in the sense that if a branch correctly describes a transitive model  $M$ , and the branch is extended by the Transitivity rule, then the resulting branch also correctly describes  $M$ . (The Transitivity rule allows adding a node  $\omega Rv$  to a branch that already contains nodes  $\omega Rv$  and  $vRv$ . If these nodes correctly describe a transitive model then so does  $\omega Rv$ .)

#### Exercise 4.6

For  $B \rightarrow C$ : If  $A$  is a conditional  $B \rightarrow C$ , then  $\beta$  contains either  $\neg B$  ( $\omega$ ) or  $C$  ( $\omega$ ). By induction hypothesis,  $M, \omega \models \neg B$  or  $M, \omega \models \neg C$ . Either way, clauses (b) and (e) of definition 3.2 imply that  $M, \omega \models A$ .

For  $\neg\Diamond B$ : If  $A$  is a negated diamond sentence  $\neg\Diamond B$ , then  $\beta$  contains a node  $\neg B$  ( $v$ ) for each world variable  $v$  for which  $\omega Rv$  is on  $\beta$  (because the tree is fully developed). By induction hypothesis,  $M, v \models \neg B$ , for each such  $v$ . By definition 4.2, it follows that  $M, v \models \neg B$  for all worlds  $v$  such that  $\omega Rv$ . By clauses (b) and (g) of definition 3.2, it follows that  $M, \omega \models A$ .

#### Exercise 4.7

We need to check that the model induced by an open branch on a fully developed K4-tree is transitive. (Suppose the model contains worlds  $w, v, u$  for which  $wRv$  and  $vRu$ . Then the Transitivity rule has been applied to the corresponding nodes on the branch, generating a node  $wRu$ . By definition 4.2,  $wRu$  holds in the induced model.)

#### Exercise 4.8

Suppose  $A$  is true at some world in some Kripke model. Then  $\neg A$  is K-invalid. Take any regular K-tree for  $\neg A$ . By observation 4.1, that tree is fully developed. By the soundness theorem for K-trees, the tree has an open branch. Let  $M$  be the model induced by some such branch  $\beta$ . Then  $M$  is acyclical. This is because the only rules that allow adding a node  $\omega Rv$  to a branch of a K-tree are the rules for expanding  $\Diamond A$  and  $\neg\Box A$  nodes. In both cases, the rule requires that the relevant world variable  $v$  is new on the branch. (Call this the *novelty requirement*.) Now suppose the accessibility relation in  $M$  has a cycle  $\omega_1 R \omega_2, \omega_2 R \omega_3, \dots, \omega_{n-1} R \omega_n, \omega_n R \omega_1$ . Each of these facts about  $R$  must correspond to a node on  $\beta$ . Of these nodes, the one that was added last (to  $\beta$ ) violates the novelty requirement. So  $M$  is acyclical.

By the Completeness Lemma, the target sentence  $\neg\neg A$  is true at world  $w$  in  $M$ . So  $A$  is true at  $w$  in  $M$ . So  $A$  is true at some world in some acyclical model.

### Exercise 4.9

The S5 rules are not sound with respect to K-validity. For example,  $\Box p \rightarrow p$  is provable with the S5 rules, but it isn't K-valid. The rules are, however, complete with respect to K-validity. This follows from the completeness of the S5 rules and the fact that every K-valid sentence is S5-valid (observation 3.1).

### Exercise 4.10

We need to show that everything that's derivable in the axiomatic calculus for S4 is true at every world in every transitive and reflexive Kripke model. From the soundness proof for K, we know that all instances of (Dual) and (K) are true at every world in every Kripke model. From observation 3.2, we know that all instances of (T) are true at every world in every reflexive Kripke model. From observation 3.3, we know that all instances of (4) are true at every world in every transitive Kripke model. So all axioms in the S4-calculus are valid in the class of transitive and reflexive Kripke frames. Since (CPL) and (Nec) preserve validity in any class of Kripke frames, it follows that everything that's derivable in the S4-calculus is valid in the class of transitive and reflexive frames.

### Exercise 4.11

(a), (b), and (c) are K-consistent, (d) is not.

### Exercise 4.12

We have to show that all S5-valid sentences are provable in the axiomatic calculus for S5, which extends the calculus for T by the axiom schemas  $\Box A \rightarrow \Box \Box A$  and  $\Diamond A \rightarrow \Box \Diamond A$ . (The second schema alone would be sufficient, as I mentioned in chapter 1, but it doesn't hurt to have the first.) The argument is by contraposition: We suppose that some sentence is not S5-provable and show that it is not S5-valid.

Suppose  $A$  is not S5-provable. Then  $\{\neg A\}$  is S5-consistent. It follows by Lindenbaum's Lemma that  $\{\neg A\}$  is included in some maximal S5-consistent set  $\Gamma$ . By definition of canonical models, this set is a world in the canonical model  $M_{S5}$  for S5. Since  $\neg A$  is in  $\Gamma$ , it follows from the Canonical Model Lemma that  $M_{S5}, \Gamma \models \neg A$ . So  $M_{S5}, S \not\models A$ .

It remains to show that the accessibility relation in  $M_{S5}$  has the right formal properties. We know that a sentence is S5-valid iff it is valid in the class of Kripke models whose accessibility relation is an equivalence relation. So we will show that the accessibility relation in  $M_{S5}$  is reflexive, transitive, and symmetric.

By definition, a world  $v$  in a canonical model is accessible from  $w$  iff whenever  $\Box A \in w$

then  $A \in v$ . Since the worlds in  $M_{S5}$  are maximal S5-consistent sets of sentences, and every such set contains every instance of the (T)-schema  $\Box A \rightarrow A$ , there is no world in  $M_{S5}$  that contains  $\Box A$  but not  $A$ . So every world in  $M_{S5}$  has access to itself.

For transitivity, suppose for some worlds  $w, v, u$  in  $M_{S5}$  we have  $wRv$  and  $vRu$ . We need to show that  $wRu$ . Given how  $R$  is defined in  $M_{S5}$ , we have to show that  $u$  contains all sentences  $A$  for which  $w$  contains  $\Box A$ . So let  $A$  be an arbitrary sentence for which  $w$  contains  $\Box A$ . Since every world in  $M_{S5}$  contains every instance of  $\Box A \rightarrow \Box \Box A$ , we know that  $w$  also contains  $\Box \Box A$ . From  $wRv$ , we can infer that  $v$  contains  $\Box A$ . And from  $vRu$ , we can infer that  $u$  contains  $A$ .

For symmetry, suppose for some worlds  $w, v$  in  $M_{S5}$  we have  $wRv$  and not  $vRw$ . Given how  $R$  is defined, this means that there is some sentence  $A$  for which  $\Box A$  is in  $v$  but  $\neg A$  is in  $w$ . Since  $w$  contains the T-provable sentence  $\neg A \rightarrow \Diamond \neg A$  and the (5)-instance  $\Diamond \neg A \rightarrow \Box \Diamond \neg A$ , it also contains  $\Box \Diamond \neg A$ . So  $v$  contains  $\Diamond \neg A$ . This contradicts the assumption that  $v$  is S5-consistent, given that  $v$  contains  $\Box A$ .

#### Exercise 4.13

(a) Method A from exercise 4.1 is sound and complete for  $X$ . (b) No set of  $\mathcal{L}_M$ -sentences is  $X$ -consistent, but every Kripke model must have at least one world.

#### Exercise 4.14

Let  $\Gamma$  be an infinite set of  $\mathcal{L}_M$ -sentences. If  $\Gamma$  is K-satisfiable then obviously every finite subset of  $\Gamma$  is satisfiable as well. For the converse direction, assume  $\Gamma$  is not K-satisfiable: There is no world in any Kripke model at which all members of  $\Gamma$  are true. Then there is no world in any Kripke model at which all members of  $\Gamma$  are true while  $p \wedge \neg p$  is false. So  $\Gamma \models p \wedge \neg p$ . By the compactness theorem, it follows that there is a finite subset  $\Gamma^-$  for which  $\Gamma^- \models p \wedge \neg p$ . If  $\Gamma^- \models p \wedge \neg p$  then there is no world in any Kripke model at which all members of  $\Gamma^-$  are true while  $p \wedge \neg p$  is false. Since  $p \wedge \neg p$  is false at every world in every Kripke model, it follows that there is no world in any Kripke model at which all members of  $\Gamma^-$  are true. This shows that if  $\Gamma$  is not K-satisfiable then there is a finite subset ( $\Gamma^-$ ) of  $\Gamma$  that is not K-satisfiable. Conversely, if every finite subset of  $\Gamma$  is K-satisfiable then  $\Gamma$  is K-satisfiable.

#### Exercise 4.15

Suppose there is a proof of  $\neg \Box(p \wedge \neg p)$ . By (CPL), we can infer  $\Box(p \wedge \neg p) \rightarrow (p \wedge \neg p)$ , because  $A \rightarrow B$  is a truth-functional consequence of  $\neg A$ . By (Nec), we get  $\Box(\Box(p \wedge \neg p) \rightarrow (p \wedge \neg p))$ .

$\neg p) \rightarrow (p \wedge \neg p)$ ). By (GL) and *modus ponens* (an instance of (CPL)), we can derive  $\Box(p \wedge \neg p)$ .

## Chapter 5

### Exercise 5.1

For an agent who knows all truths only the actual world is epistemically accessible. For an agent who knows nothing all worlds are epistemically accessible.

### Exercise 5.2

- (a)  $K(r \vee s)$   
 $r$ : It is raining;  $s$ : It is snowing
- (b)  $K r \vee K s$   
 $r$ : It is raining;  $s$ : It is snowing
- (c)  $K r \vee K \neg r$   
 $r$ : It is raining
- (d) This sentence is ambiguous. On one reading, it could be translated as  $M g \rightarrow K g$ , on the other as  $K(M g \rightarrow g)$   
 $g$ : You are guilty

### Exercise 5.3

You can use [umsu.de/trees/](https://umsu.de/trees/) to create the tree proof. We can assume S5 for the box because it quantifies unrestrictedly over all worlds (as in chapter 2).

### Exercise 5.4

(NT) is valid on all and only the frames in which no world can see any world.

### Exercise 5.5


We assume that ignorance of  $A$  can be formalized as  $A \wedge \neg K A$ . Ignorance of ignorance of  $A$  is therefore formalized as  $(A \wedge \neg K A) \wedge \neg K(A \wedge \neg K A)$ . A tree proof shows that the former  $K$ -entails the latter.

### Exercise 5.6

In a Gettier case, the relevant proposition  $p$  (say, that you're looking at a barn) is true but unknown. By (0.4), it would follow that the agent knows that they don't know  $p$ . But in a typically Gettier case the agent does not know that they don't know  $p$ .

**Exercise 5.7**

All except (a) and (d) are correct. You can find trees or counterexamples for (a)-(e) on [umsu.de/trees/](https://umsu.de/trees/) if you write K as a box and M as a diamond. Here is a tree for (f):

1.	$\neg((M K p \wedge M K q) \rightarrow M K(p \wedge q))$	(w)	(Ass.)
2.	$M K p \wedge M K q$	(w)	(1)
3.	$\neg M K(p \wedge q)$	(w)	(1)
4.	$M K p$	(w)	(2)
5.	$M K q$	(w)	(2)
6.	$w R v$		(4)
7.	$K p$	(v)	(4)
8.	$w R u$		(5)
9.	$K q$	(u)	(5)
10.	$v R t$		(6,8,Con)
11.	$u R t$		(6.8,Con)
12.	$w R t$		(6.10,Tr)
13.	$\neg K(p \wedge q)$	(t)	(3,12)
14.	$t R s$		(13)
15.	$\neg(p \wedge q)$	(s)	(13)
			
16.	$\neg p$	(s)	(15)
17.	$\neg q$	(s)	(15)
18.	$v R s$		(10.14,Tr)
19.	$u R s$		(11.14,Tr)
20.	$p$	(s)	(7,18)
21.	$q$	(s)	(9,19)
	x		x

**Exercise 5.8**

see <https://plato.stanford.edu/entries/dynamic-epistemic/appendix-B-solutions.html> (where all the dates are 10 days later than they are in my version).

**Exercise 5.9**

(a) and (b) are valid, (c) and (d) are invalid. Here is a tree proof for (a).

1.  $\neg(M_1 K_2 p \rightarrow M_1 p)$  (w) (Ass.)
2.  $M_1 K_2 p$  (w) (1)
3.  $\neg M_1 p$  (w) (1)
4.  $wR_1 v$  (2)
5.  $K_2 p$  (v) (2)
6.  $\neg p$  (v) (3,4)
7.  $vR_2 v$  (Refl.)
8.  $p$  (v) (5,7)
- x

The tree for (c) doesn't close:

1.  $\neg(M_1 K_2 p \rightarrow M_2 K_1 p)$  (w) (Ass.)
2.  $M_1 K_2 p$  (w) (1)
3.  $\neg M_2 K_1 p$  (w) (1)
4.  $wR_1 v$  (2)
5.  $K_2 p$  (v) (2)
6.  $vR_2 v$  (Refl.)
7.  $p$  (v) (5,6)
8.  $wR_2 w$  (Refl.)
9.  $\neg K_1 p$  (w) (3,8)
10.  $wR_1 u$  (9)
11.  $\neg p$  (u) (9)

We could add a few more applications of Reflexivity, but the tree would remain open. It also gives us a countermodel: let  $W = \{w, v, u\}$ ;  $w$  has 1-access to  $v$  and  $u$ ; each world has 1- and 2-access to itself;  $V(p) = \{v\}$ . In this model, at world  $w$ ,  $M_1 K_2 p$  is true while  $M_2 K_1 p$  is false.

Cases (b) and (d) are similar.

#### Exercise 5.10

The (5)-schema for  $E_G$  states that  $\neg E_G \neg A \rightarrow E_G \neg E_G \neg A$ . To show that some instance of this is invalid, we need to find a case where some instance of  $\neg E_G \neg A$  is true while

$E_G \neg E_G \neg A$  is false. We can take the simplest instance, with  $A = p$ . Assume the relevant group has two agents, and consider a world  $w$  at which  $K_1 \neg p$  and  $\neg K_2 \neg p$  are true. By the assumption that (5) is valid for  $K_i$ ,  $K_2 \neg K_2 \neg p$  is also true at  $w$ . But  $K_1 \neg K_2 \neg p$  can be false (at  $w$ ). If it is, then  $\neg E_G \neg p$  is true at  $w$  while  $E_G \neg E_G \neg p$  is false.

#### Exercise 5.11

No, a transitive, serial, and euclidean relation is not always symmetric. Counterexample:  $wRv, vRv$ . This means that not all instances of (B) (which corresponds to symmetry) are valid in KD45.

#### Exercise 5.12

You can e.g. do a tree proof, using B as the box.

#### Exercise 5.13

Let  $A$  be an arbitrary proposition.

By (PI),  $BA \rightarrow KBA$  is valid. By (KB), so is  $KBA \rightarrow BBA$ . By propositional logic, these entail  $BA \rightarrow BBA$ .

By (NI),  $\neg B \neg A \rightarrow K \neg B \neg A$  is valid. By (KB), so is  $K \neg B \neg A \rightarrow B \neg B \neg A$ . By propositional logic, these entail  $\neg B \neg A \rightarrow B \neg B \neg A$ .

#### Exercise 5.14

The left-to-right direction is (KB). For the right-to-left direction, let  $A$  be an arbitrary proposition. By (SB),  $BA \rightarrow BKA$  is valid. By (D) for belief,  $BKA \rightarrow \neg B \neg KA$  is valid. The contraposition of (KB) gives us  $\neg B \neg KA \rightarrow \neg K \neg KA$ . Finally, the contraposition of (5) for knowledge yields  $\neg K \neg A \rightarrow KA$ . The target proposition  $BA \rightarrow KA$  is a truth-functional consequence of these four propositions.

#### Exercise 5.15

If the logic of belief is KD45 then  $\Box \Diamond p$  is equivalent to  $\Diamond p$  (as you can show, for example, with a tree proof).

#### Exercise 5.16

Suppose  $B(p \wedge \neg Bp)$ . In any logic that extends K, it follows that  $Bp$  and  $B \neg Bp$ . By (4),  $Bp$  entails  $B Bp$ . Now we have  $B \neg Bp$  and  $B Bp$ , which violates (D).

## Chapter 6

### Exercise 6.1

- (a)  $O \neg p$ ;  $p$ : You go into the garden.
- (b)  $O \neg p$ ;  $p$ : You go into the garden.
- (c)  $O p$ ;  $p$ : Jones helps his neighbours.
- (d)  $O(p \rightarrow q)$ ;  $p$ : Jones helps his neighbours,  $q$ : Jones tells his neighbours that he's coming.
- (e) You might try  $O(\neg p \rightarrow \neg q)$  or  $\neg p \rightarrow O \neg q$   $p$ : Jones helps his neighbours,  $q$ : Jones tells his neighbours that he's coming.

See section 6.3, especially exercise 6.14, for why neither translation of (e) is fully satisfactory.

### Exercise 6.2

No, not unless all (or no) worlds are ideal. If  $w$  is ideal and  $v$  is not, then  $wRv$  but not  $vRw$ .

### Exercise 6.3

$R$  is euclidean if  $\forall x \forall y \forall z ((xRy \wedge xRz) \rightarrow yRz)$ . Suppose  $wRv$ . Instantiating the universal formula with  $w$  for  $x$  and with  $v$  for  $y$  and  $z$ , we have  $(wRv \wedge wRv) \rightarrow vRv$ . So  $vRv$ .

### Exercise 6.4

Let  $w$  be a world in which the only relevant norm is that one must drive on the left. Let  $v$  and  $u$  be worlds in which everyone drives on the left even though the law requires driving on the right. Both  $v$  and  $u$  are accessible from  $w$ , but  $u$  is not accessible from  $v$ , and  $v$  is not accessible from itself.

### Exercise 6.5

Use <https://www.umsu.de/trees/>. (Write  $O$  as a box and  $P$  as a diamond. For  $D$ , make the accessibility relation serial; for  $KD45$ , make it serial, transitive, and euclidean.)

### Exercise 6.6

(Dual1) says that  $\neg \Diamond A$  is equivalent to  $\Box \neg A$ . If nothing is permitted then  $\neg \Diamond A$  is true for all  $A$ . But if nothing is forbidden then  $\Box \neg A$  is false for all  $A$ .

(Dual2) says that  $\neg\Box A$  is equivalent to  $\Diamond\neg A$ . If nothing is forbidden then  $\neg\Box A$  is true for all  $A$ . But if nothing is permitted then  $\Diamond\neg A$  is false for all  $A$ .

### Exercise 6.7

We assume that every world has some norms. (See the previous exercise.) Suppose a world  $v$  is accessible from a world  $w$ , meaning that  $v$  conforms to the norms of  $w$ . Since the norms are intolerant, the norms at  $v$  must be the same as at  $w$ . So whatever is accessible from  $v$  is accessible from  $w$ . It follows that the accessibility relation is transitive and euclidean. (In fact, the complete relativist logic of intolerant norms is KD45.)

Note that the absolutist approach validates schemas like  $OA \rightarrow OOA$  *without* assuming that the norms say anything about whether they should be in force.

### Exercise 6.8

$PA$  could be defined as  $\neg\Box(N \rightarrow \neg A)$ , or more simply (and equivalently) as  $\Diamond(N \wedge A)$ .

### Exercise 6.9

In the Leibnizian language, the (U)-schema turns into  $\Box(N \rightarrow (\Box(N \rightarrow A) \rightarrow A))$ . You can use a tree proof to show that this is T-valid. (See [umsu.de/trees/](http://umsu.de/trees/).)

### Exercise 6.10

In the described situation, it ought to be the case that Amy is either obligated to help Betty or obligated to help Carla, but Amy is neither obligated to help Betty nor to help Carla. So if  $p$  translates ‘Amy helps Betty’ and  $q$  ‘Amy helps Carla’, we seem to have  $O(Op \vee Oq)$  and  $\neg Op$  and  $\neg Oq$ . But these assumptions are inconsistent in K5. You can draw a K5-tree (using the K-rules and the Euclidity rule) starting with  $O(Op \vee Oq)$  and  $\neg Op$  and  $\neg Oq$  on which all branches close. This shows that there is no world in any euclidean model at which the three assumptions are true.

### Exercise 6.11

Since we assume that there is always at least one best world among the accessible worlds, and the accessible worlds comprise just one world, it follows that  $OA$  is true at  $w$  iff  $A$  is true at  $w$ . The logic we get is the “Triv” logic that is axiomatized by adding the (Triv)-schema  $\Box A \leftrightarrow A$  to the standard axioms and rules for K. This logic is stronger than S5:

all S5-valid sentences are Triv-valid. We also have, among other things, all instances of  $\Box A \leftrightarrow \Diamond A$ . The choice between absolutism and relativism makes no difference.

#### Exercise 6.12

Use [umsu.de/trees/](http://umsu.de/trees/).

#### Exercise 6.13

Deontic detachment is valid. 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  $B$  is true at the best of the accessible worlds.

Factual detachment is invalid. A counterexample is 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  $O(g/k)$ , but not  $O(g)$ . Formally,  $g$  is true at the best of the accessible  $k$ -worlds, but since all the  $k$ -worlds are quite bad,  $g$  is not true at the best of the accessible worlds.

#### Exercise 6.14

(c) can obviously be translated as  $O p$ , (f) as  $\neg p$ .

You probably translated (d) as either  $p \rightarrow O q$  or as  $O(p \rightarrow q)$ .  $p \rightarrow O q$  is entailed by (f), so it violates the non-entailment condition. Assume then that (d) is translated as  $O(p \rightarrow q)$ .

The most obvious translations for (e) are  $\neg p \rightarrow O \neg q$  and  $O(\neg p \rightarrow \neg q)$ . The latter is entailed by (c). If we choose the former, then (c)–(f) constitute a deontic dilemma: (e) and (f) would entail  $O \neg t$ , but (c) and (d) would entail  $O t$ .

#### Exercise 6.15

Simply replace ‘all’ in the semantics for  $O(B/A)$  with ‘some’.

#### Exercise 6.16

Ross’s Paradox: ‘Alice must be in the office or in the library’ seems to imply that Alice might be in the office and that she might be in the library.

The Paradox of Free Choice: ‘Alice might be in the office or in the library’ seems to imply that Alice might be in the office and that she might be in the library.

**Exercise 6.17**

For every world  $w$ , every member of  $N(w)$  contains  $w$ .

**Exercise 6.18**

In Kripke semantics,  $\Box p$  and  $\Box q$  together entail  $\Box(p \wedge q)$ . But if the probability of  $p$  is above the threshold and the probability of  $q$  is above the threshold, it does not follow that the probability of  $p \wedge q$  is above the threshold. For example, we could have  $\Pr(p) = 0.95$ ,  $\Pr(q) = 0.94$ , and  $\Pr(p \wedge q) = 0.95 \times 0.94 = 0.893$ .

## Chapter 7

### Exercise 7.1

- (a)  $H \neg p$   
 $p$ : It is warm
- (b)  $F p$   
 $p$ : There is a sea battle
- (c)  $\neg F P p$  or, perhaps,  $F \neg P p$   
 $p$ : There is a sea battle
- (d)  $F(p \vee P q)$  or  $F(F p \vee F P q)$   
 $p$ : It is warm
- (e)  $\neg P p \rightarrow \neg F q$  or  $G(\neg P p \rightarrow \neg q)$   
 $p$ : You study,  $q$ : you pass the exam
- (f)  $P(p \wedge q)$   
 $p$ : I am having tea,  $q$ : the door bell rings

### Exercise 7.2

(a), (c), (f), (g), and (h) are true, (b), (d), and (e) are false.

### Exercise 7.3

(Con1): Suppose some sentence of the form  $A \rightarrow G P A$  is false at some time  $t$  in some temporal model. By clause (e) of definition 7.2, this means that  $A$  is true at  $t$  and  $G P A$  is false at  $t$ . By clause (h), the latter means that there is a time  $s$  with  $t < s$  such that  $P A$  is not true at  $s$ . By clause (i), it follows that  $A$  is not true at  $t$ . Contradiction.

The argument for (Con2) is analogous.

### Exercise 7.4

- (a) 1.  $\neg(A \rightarrow GPA)$  (t) (Ass.)  
 2.  $A$  (t) (1)  
 3.  $\neg GPA$  (t) (1)  
 4.  $t < s$  (3)  
 5.  $\neg PA$  (s) (3)  
 6.  $\neg A$  (t) (4,5)  
 x
- (b) 1.  $\neg(A \rightarrow HFA)$  (t) (Ass.)  
 2.  $A$  (t) (1)  
 3.  $\neg HFA$  (t) (1)  
 4.  $s < t$  (3)  
 5.  $\neg FA$  (s) (3)  
 6.  $\neg A$  (t) (4,5)  
 x
- (c) 1.  $\neg(FA \rightarrow HFFA)$  (t) (Ass.)  
 2.  $FA$  (t) (1)  
 3.  $\neg HFFA$  (t) (1)  
 4.  $s < t$  (3)  
 5.  $\neg FFA$  (s) (3)  
 6.  $\neg FA$  (t) (4,5)  
 x
- (d) 1.  $\neg(PGA \rightarrow PFA)$  (t) (Ass.)  
 2.  $PGA$  (t) (1)  
 3.  $\neg PFA$  (t) (1)  
 4.  $s < t$  (2)  
 5.  $GA$  (s) (2)  
 6.  $A$  (t) (4,5)  
 7.  $\neg FA$  (s) (3,4)  
 8.  $\neg A$  (t) (4,7)  
 x

(e)	1.	$\neg(HA \leftrightarrow HFHA)$	(t)	(Ass.)	
2.	$HA$	(t)	(1)	4.	$\neg HA$ (t) (1)
3.	$\neg HFHA$	(t)	(1)	5.	$HFHA$ (t) (1)
6.	$s < t$		(3)	9.	$s < t$ (4)
7.	$\neg FHA$	(s)	(3)	15.	$\neg A$ (s) (4)
8.	$\neg HA$	(t)	(6,7)	11.	$FHA$ (s) (5,9)
	x			12.	$s < r$ (11)
				13.	$HA$ (r) (11)
				14.	$A$ (s) (12,13)
					x

#### Exercise 7.5

Suppose  $<$  is transitive, and  $x > y$  and  $y > z$ . Equivalently,  $y < x$  and  $z < y$ . By transitivity of  $<$ , we have  $z < x$ . So  $x > z$ .

#### Exercise 7.6

Suppose  $R$  is transitive. If there are points  $x$  and  $y$  for which  $xRy$  and  $yRx$  then  $xRx$  by transitivity. So if  $R$  isn't asymmetric then it isn't irreflexive. If  $R$  isn't irreflexive then there is a point  $x$  with  $xRx$ . This violates asymmetry, because asymmetry demands that if  $xRx$  then not  $xRx$ .

#### Exercise 7.7

If time is transitive and circular, then it is neither asymmetric nor irreflexive.

#### Exercise 7.8

(a), (d), and (e) are invalid. Here are trees for (b), (c), and (f). I can't typeset the one for (g).

- (b)
- |    |   |     |        |
|----|---|-----|--------|
| 1. | $\neg(P \ G \ G p \rightarrow G \ G p)$ | (t) | (Ass.) |
| 2. | $P \ G \ G p$                           | (t) | (1)    |
| 3. | $\neg \ G \ G p$                        | (t) | (1)    |
| 4. | $s < t$                                 |     | (2)    |
| 5. | $G \ G p$                               | (s) | (2)    |
| 6. | $t < r$                                 |     | (3)    |
| 7. | $\neg \ G p$                            | (r) | (3)    |
| 8. | $s < r$                                 |     | (3,6)  |
| 9. | $G p$                                   | (r) | (5,8)  |
|    | x                                       |     |        |

- (c)
- |     |   |     |        |
|-----|---|-----|--------|
| 1.  | $\neg(P \ F p \rightarrow (P p \vee (p \vee F p)))$ | (t) | (Ass.) |
| 2.  | $P \ F p$   | (t) | (1)    |
| 3.  | $\neg(P p \vee (p \vee F p))$                       | (t) | (1)    |
| 4.  | $\neg P p$  | (t) | (3)    |
| 5.  | $\neg(p \vee F p)$                                  | (t) | (3)    |
| 6.  | $\neg p$  | (t) | (5)    |
| 7.  | $\neg F p$  | (t) | (5)    |
| 8.  | $s < t$   |     | (2)    |
| 9.  | $F p$   | (s) | (2)    |
| 10. | $s < r$   |     | (9)    |
| 11. | $p$   | (r) | (9)    |
- 
- |     |          |     |        |     |          |     |        |     |          |     |        |
|-----|----------|-----|--------|-----|----------|-----|--------|-----|----------|-----|--------|
| 12. | $t < r$  |     |        | 13. | $t = r$  |     |        | 14. | $r < t$  |     |        |
| 15. | $\neg p$ | (r) | (7,12) | 16. | $\neg p$ | (r) | (6,13) | 17. | $\neg p$ | (r) | (4,16) |
|     | x        |     |        |     | x        |     |        |     | x        |     |        |

(f)	1.	$\neg(F(Gq \wedge \neg p) \rightarrow G(p \rightarrow (Gp \rightarrow q)))$	(t)	(Ass.)
	2.	$F(Gq \wedge \neg p)$	(t)	(1)
	3.	$\neg G(p \rightarrow (Gp \rightarrow q))$	(t)	(1)
	4.	$t < s$		(2)
	5.	$Gq \wedge \neg p$	(s)	(2)
	6.	$Gq$	(s)	(5)
	7.	$\neg p$	(s)	(5)
	8.	$t < r$		(3)
	9.	$\neg(p \rightarrow (Gp \rightarrow q))$	(r)	(3)
	10.	$p$	(r)	(9)
	11.	$\neg(Gp \rightarrow q)$	(r)	(9)
	12.	$Gp$	(r)	(11)
	13.	$\neg q$	(r)	(11)
	14.	$s < r$		
	15.	$s = r$		
	16.	$r < s$		
	17.	$q$	(r)	(6,14)
		x		
	18.	$p$	(s)	(10,15)
		x		
	19.	$p$	(s)	(12,16)
		x		

### Exercise 7.9

- (a) For example,  $GA \rightarrow FA$ .
- (b) For example,  $HA \rightarrow PA$ .
- (c) No schema corresponds to the class of frames with a last time. If we also assume linearity,  $G(A \wedge \neg A) \vee FG(A \wedge \neg A)$  works.
- (d) No schema corresponds to the class of frames with a first time. If we assume linearity, then  $H(A \wedge \neg A) \vee PH(A \wedge \neg A)$  works.

### Exercise 7.10

Assume a frame is dense. Suppose for reductio that some instance of  $FA \rightarrow FFA$  is false at some point  $t$  in some model  $M$  based on that frame. Then  $FA$  is true at  $t$  and  $FFA$  is false. Since  $FA$  is true at  $t$ , it follows by definition 7.2 that  $A$  is true at some point  $s$  such that  $t < s$ . By density, there is a point  $r$  such that  $t < r < s$ . But since  $A$  is true at  $s$ ,  $FA$  is true at  $r$ , and so  $FFA$  is true at  $t$ ; contradiction.

In the other direction, we have to show that if a frame isn't dense then some instance of  $\text{FA} \rightarrow \text{FFA}$  is false at some point  $t$  in some model  $M$  based on that frame. We take the simplest instance  $\text{Fp} \rightarrow \text{FFp}$ . If a frame isn't dense then there are points  $t, s$  such that  $t < s$  and no point lies in between  $t$  and  $s$ . Let  $V$  be an interpretation function that makes  $p$  true at  $s$  and false everywhere else. Then  $\text{Fp}$  is true at  $t$  but  $\text{FFp}$  is false. So  $\text{Fp} \rightarrow \text{FFp}$  is false at  $t$ .

### Exercise 7.11

Without assumptions about the flow of time there is no way to express in  $\mathcal{L}_T$  that  $p$  is true at all times (or at some time). In linear flows,  $p \wedge \text{Hp} \wedge \text{Gp}$  does the job.

### Exercise 7.12

Consider a model with three times ordered by  $s < t$  and  $s < r$ . Assume  $p$  is true at  $t$  and not at  $r$ . Then  $p \rightarrow \text{HFp}$  is false on the Peircean interpretation.

### Exercise 7.13

(a)–(d) are valid, (e) is invalid.

To show that a schema is valid, assume for reductio that there is some time  $t$  on some history  $H$  in some model  $M$  at which the schema is false. Then (repeatedly) use definition 7.3 to derive a contradiction.

For (e), consider a model with three times  $t, s, r$  such that  $s < t, r < t$ , and neither  $s < r$  nor  $r < s$ . Let  $q$  be true at  $s$  and false at the other two times.  $\text{P}q \rightarrow \Box \text{P} \Diamond q$  is false at  $t$  on the history  $\langle s, t \rangle$ .

### Exercise 7.14

A sentence  $A$  is super-valid iff  $M, t \models A$  for all temporal models  $M$  and times  $t$  in  $M$ . By supervaluationism, this holds iff  $M, H, t \models A$  for all  $M, t$ , and histories  $H$  containing  $t$ . That's how Ockhamist validity was originally defined.

### Exercise 7.15

Ockham-entailment is stronger than super-entailment: whenever  $A$  Ockham-entails  $B$ , then  $A$  super-entails  $B$ , but not the other way around.

Suppose  $A$  Ockham-entails  $B$ . Let  $t$  be any time in any temporal model at which  $A$  is true, i.e.: true relative to all histories through  $t$ . Since  $A$  Ockham-entails  $B$ ,  $B$  is true at  $t$

relative to all histories through  $t$ . So  $A$  super-entails  $B$ .

But suppose  $A$  super-entails  $B$ . Let  $t$  be any time on any history  $h$  in any temporal model at which  $A$  is true. We can't infer that  $B$  is true at  $t$  on  $h$ , for  $A$  may be false at  $t$  relative to other histories  $h'$ . So we can't infer that  $A$  Ockham-entails  $B$ . Indeed,  $\text{F}p$  super-entails  $\Box \text{F}p$ , but  $\text{F}p$  does not Ockham-entail  $\Box \text{F}p$ .

**Exercise 7.16**

$(A \wedge \neg A) \cup A$ .

**Exercise 7.17**

$Ap \rightarrow p$ .

## Chapter 8

### Exercise 8.1

(E1)–(E5) are invalid assuming that ‘if  $A$  then  $B$ ’ is true iff both  $A$  and  $B$  are true. There are, of course, strong reasons against the analysis of English conditionals as conjunctions.

### Exercise 8.2

For example:  $\neg A \rightarrow A$  or  $(A \vee \neg A) \rightarrow A$ .

### Exercise 8.3

$W = \{w\}$ ,  $R = \emptyset$ ,  $V(p) = \{w\}$ ,  $V(q) = \emptyset$ .

### Exercise 8.4

Use [umsu.de/trees/](http://umsu.de/trees/).

### Exercise 8.5

(E1)–(E5) all work equally well in the subjunctive mood. For (E4) and (E5):

- If our opponents had been cheating, we would never have found out. Therefore: If we had found out that our opponents are cheating, then they wouldn’t have been cheating.
- If you had added sugar to your coffee, it would have tasted good. Therefore: If you had added sugar and vinegar to your coffee, it would have tasted good.

Both of these inferences are valid if subjunctive conditionals are strict conditionals. But they don’t sound good.

### Exercise 8.6

Suppose  $A \rightarrow B$  is assertable. Then  $A \rightarrow B$  is known. So  $K(A \rightarrow B)$ . In S4, it follows that  $KK(A \rightarrow B)$ . So the epistemically strict conditional  $K(A \rightarrow B)$  is assertable. Conversely, if  $K(A \rightarrow B)$  is assertable, then it is known; so  $KK(A \rightarrow B)$ . In S4, it follows that  $K(A \rightarrow B)$ . So  $A \rightarrow B$  is assertable.

### Exercise 8.7

The ‘or-to-if’ inference is not valid on the assumption that the conditional is epistemically strict. For example, if  $p$  and  $q$  are both true at the actual world and both false at some

epistemically accessible world, then ‘ $p$  or  $q$ ’ is true but ‘if  $p$  then  $q$ ’ is false (on the strict analysis).

The inference might nonetheless look reasonable because it would normally be inappropriate to assert a disjunction ‘ $p$  or  $q$ ’ unless the disjunction is known – unless it is true at all epistemically accessible worlds. And if  $p \vee q$  is true at all epistemically accessible worlds then  $\neg p \rightarrow q$  is also true at all epistemically accessible worlds, and so  $\Box(\neg p \rightarrow q)$  is true. Thus the conclusion of or-to-if is true in any situation in which the premise is *assertable*. If the logic of knowledge validates the (4)-schema, we can go further and say that the conclusion is assertable in any situation in which the premise is assertable.

### Exercise 8.8

Assume that  $R$  is asymmetric and quasi-connected. We want to show that  $R$  is transitive. So assume we have  $xRy$  and  $yRz$ . By quasi-connectedness,  $yRz$  implies that either  $yRx$  or  $xRz$ . By asymmetry, we can’t have  $yRx$ , since we have  $xRy$ . So  $xRz$ .

### Exercise 8.9

We have the following equivalences (using ‘ $\Leftrightarrow$ ’ to mean that the expressions on either side are equivalent):

$$u \not\leq_w v \Leftrightarrow \neg(u \leq_w v) \Leftrightarrow \neg(v \not\leq_w u) \Leftrightarrow v <_w u.$$

So you can simply replace every instance of  $\omega <_w v$  in the conditions by  $v \not\leq_w \omega$ , and every instance of  $\omega \not\leq_w v$  by  $v \leq_w \omega$ .

Asymmetry thereby turns into: if  $u \not\leq_w v$  then  $v \leq_w u$ . Equivalently: either  $u \leq_w v$  or  $v \leq_w u$ . This property of relations is called **completeness**. Notice that it entails reflexivity.

Quasi-connectedness turns into: if  $u \not\leq_w v$  then for all  $t$ , either  $t \not\leq_w v$  or  $u \not\leq_w t$ . This is equivalent to transitivity for  $\leq$ .

The Limit Assumption turns into: for any non-empty set of worlds  $X$  and world  $w$  there is a  $v \in X$  such that there is no  $u \in X$  with  $v \not\leq_w u$ . Equivalently, for any non-empty set of worlds  $X$  and world  $w$  there is a  $v \in X$  such that  $v \leq_w u$  for all  $u \in X$ .

### Exercise 8.10

No. We don’t want  $A$  and  $O(B/A)$  to entail  $B$ . Semantically, we don’t want to assume that every world is among the best worlds relative to its own norms.

### Exercise 8.11

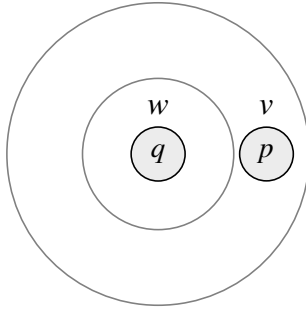
Suppose  $A \Box \rightarrow B$  is true at some world  $w$  in some model  $M$ . So  $B$  is true at all the closest  $A$ -worlds to  $w$ . Now either  $A$  is true at  $w$  or  $A$  is false at  $w$ . If  $A$  is false at  $w$ , then  $A \rightarrow B$  is true at  $w$ . If  $A$  is true at  $w$ , then  $w$  is one of the closest  $A$ -worlds to  $w$ , by Weak Centring; so  $B$  is true at  $w$ ; and so  $A \rightarrow B$  is true at  $w$ . Either way, then,  $A \rightarrow B$  is true at  $w$ .

### Exercise 8.12

If  $A$  is true at no worlds, then  $\text{Min}^{<_w}(\{u : M, u \models A\})$  is the empty set. So it is vacuously true that  $M, v \models B$  for all  $v \in \text{Min}^{<_w}(\{u : M, u \models A\})$ .

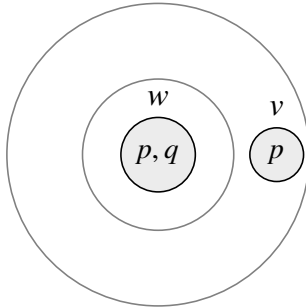
### Exercise 8.13

(E1) is an inference from  $q$  to  $p \Box \rightarrow q$ . To show that this is invalid, we need to give a model in which  $q$  is true at some world ( $w$ ) while  $p \Box \rightarrow q$  is false (at  $w$ ).



This model also shows that (E2) and (E3) are invalid. (E2) is an inference from  $\neg p$  to  $p \Box \rightarrow q$ . In the model,  $\neg p$  is true at  $w$  but  $p \Box \rightarrow q$  is false. (E3) is an inference from  $\neg(p \Box \rightarrow q)$  to  $p$ . In the model,  $\neg(p \Box \rightarrow q)$  is true at  $w$  but  $p$  is false.

(E4) is an inference from  $p \Box \rightarrow q$  to  $\neg q \Box \rightarrow \neg p$ . In the following model, the premise is true at  $w$  and the conclusion false.



### Exercise 8.14

Frances has never learnt a foreign language, although she would have loved to learn French. If Frances had been given a choice between learning French and learning Italian, she would have chosen French. *If Frances had learned French or Italian then she would have learned French.* It does not follow that if Frances had learned Italian then she would have learned French.

The same style of example works for indicative conditionals.

### Exercise 8.15

- (a) Assume  $A \wedge B$  is true at some world  $w$  in some model  $M$ . By Centring,  $w$  is among the closest  $A$ -worlds to  $w$ . By connectedness,  $w$  is the unique closest  $A$ -world to  $w$ . So  $B$  is true at all closest  $A$ -worlds to  $w$ .
- (b) Assume  $A \Box \rightarrow (B \vee C)$  is true at some world  $w$  in some model  $M$ . So all the closest  $A$ -worlds to  $w$  are  $(B \vee C)$ -worlds. If there are no  $A$ -worlds then  $A \Box \rightarrow B$  and  $A \Box \rightarrow C$  are both true. If there are  $A$ -worlds then Stalnaker's semantics implies that there is a unique closest  $A$ -world  $v$  to  $w$ . Since  $B \vee C$  is true at  $v$ , either  $B$  or  $C$  must be true at  $v$ . So either  $B$  is true at all closest  $A$ -worlds to  $w$  or  $C$  is true at all closest  $A$ -worlds to  $w$ .

### Exercise 8.16

'All dogs are barking':  $\forall x(Dx \rightarrow Bx)$

'Some dogs are barking':  $\exists x(Dx \wedge Bx)$

'Most dogs are barking' cannot be translated in terms of  $Mx$ . We need a binary quantifier:  $Mx(Bx/Dx)$

### Exercise 8.17

On this proposal, bare indicative conditionals like (8) are material conditionals. If  $p$  is true and  $q$  is false then there is an accessible  $p$ -world at which  $q$  is false, and so  $q$  is not true at all accessible worlds at which  $p$  is true. In all other cases,  $q$  is true at all accessible worlds at which  $p$  is true.

### Exercise 8.18

Conditional Excluded Middle is valid iff there is never more than one closest/accessible  $A$ -world. On that assumption, 'some closest/accessible  $A$ -world is a  $B$ -world' entails 'all closest/accessible  $A$ -worlds are  $B$ -worlds'. But (10) does not entail 'If I had played the lottery, I would have won'.

## Chapter 9

### Exercise 9.1

- (a)  $Srj \wedge Skj$ ;  $r$ : Keren,  $k$ : Keziah,  $j$ : Jemima,  $S$ : – is a sister of –
- (b)  $\forall x(Mx \rightarrow Ox)$ ;  $M$ : – is a myriapod,  $O$ : – is oviparous
- (c)  $\exists x(Cx \wedge Nx \wedge Hfx)$ ;  $f$ : Fred,  $C$ : – is a car,  $N$ : – is new,  $H$ : – has –
- (d)  $\neg \forall x(Sx \rightarrow Lxl)$ ;  $l$ : logic;  $S$ : – is a student,  $L$ : – loves –
- (e)  $\forall x((Sx \wedge Lxl) \rightarrow \exists yLxy)$ ;  $l$ : logic;  $S$ : – is a student,  $L$ : – loves –

### Exercise 9.2

Let the model  $M$  be given by  $D = \{\text{Rome}, \text{Paris}\}$  and  $V(F) = \{\text{Rome}\}$ . By clause (a) of definition 9.2,  $M, g' \models Fx$  holds for every assignment function  $g'$  that maps  $x$  to Rome, because then  $g'(x) \in V(F)$ . By clause (h) it follows that  $M, g \models \exists xFx$  for every assignment function  $g$ . By clause (a) again,  $M, g' \not\models Fx$  for every assignment function  $g'$  that maps  $x$  to Paris. By clause (g), it follows that  $M, g \not\models \forall xFx$  for every assignment function  $g$ . So  $\exists xFx$  is true (in  $M$ ) relative to every assignment function while  $\forall xFx$  is false relative to every assignment function. By clause (e) it follows that  $\exists xFx \rightarrow \forall xFx$  is false in  $M$  relative to every assignment function.

### Exercise 9.3

For both cases, use  $Fx$  as the sentence  $A$ , and  $\neg Fx$  as  $B$ , and consider a model in which  $F$  applies to some but not to all individuals. Both  $Fx$  and  $\neg Fx$  are then true relative to some assignment functions and false relative to others. So neither sentence is true in the model. But  $Fx \vee \neg Fx$  is true relative to every assignment function.

### Exercise 9.4

(a) There are many non-reflexive models in which  $\Box p \rightarrow p$  is true at some world – for example, any non-reflexive model in which  $p$  is false at all worlds.

(b) Let  $M_1$  be a model with a single world that can see itself. Let  $M_2$  be a model with two worlds, each of which can see the other but not itself. In both models, all sentence letters are false at all worlds. The very same  $\mathcal{L}_M$ -sentences are true at all worlds in these models (as a simple proof by induction shows). But the first model is reflexive and the second isn't. So there is no  $\mathcal{L}_M$ -question that is true at a world in a model iff the model's accessibility relation is reflexive.

### Exercise 9.5

Use [umsu.de/trees/](http://umsu.de/trees/).

### Exercise 9.6

If a sentence is valid (in first-order predicate logic) then a fully expanded tree for the sentence will close and show that the sentence is valid. But if a sentence is not valid, the tree might grow forever. There is no algorithm for detecting whether a tree will grow forever.

### Exercise 9.7

- (a)  $\Box Fa$

$a$ : John,  $F$ : – is hungry.

(Might be classified as either *de re* or *de dicto*.)

- (b)  $\Box \forall x(Fx \rightarrow Gx)$

$F$ : – is a cyclist,  $G$ : – has legs.

This is *de dicto*. Also correct (but different in meaning) is the *de re* translation  $\forall x(Fx \rightarrow \Box Gx)$ . Close but incorrect (and *de re*):  $\forall x\Box(Fx \rightarrow Gx)$ .

- (c)  $\forall x(Fx \rightarrow \Diamond Gx)$

$F$ : – is a day,  $G$ : – is our last day.

Better:  $\forall x(Fx \rightarrow \Diamond(Hx \wedge \neg \exists y(Fy \wedge Lyx \wedge Hy)))$

$F$ : – is a day,  $L$ : – is later than –,  $H$ : We are alive on –.

Both *de re*. The English sentence could also be understood *de dicto*, as  $\Diamond \forall x(Fx \rightarrow Gx)$ , but that would be a very strange thing to say.

- (d) I would translate this as  $\forall x O(Fx \rightarrow Gx)$

$F$ : – wants to leave early,  $G$ : – leaves quietly.

Even better, if we can use the conditional obligation operator:  $\forall x O(Gx/Fx)$ . Also defensible are  $\forall x(Fx \rightarrow O Gx)$  and  $O \forall x(Fx \rightarrow Gx)$ .

All of these are *de re*.

- (e)  $\forall x(\exists y(Fy \wedge Hxy) \rightarrow P Gx)$

$F$ : – is a ticket,  $G$ : – enters,  $H$ : – bought –.

Perhaps even better:  $\forall x P(Gx/\exists y(Fy \wedge Hxy))$ . Both of these are *de re*.

You could translate ‘bought a ticket’ as a simple predicate here; you could also use

a temporal operator to account for the past tense of ‘bought’ (but it’s confusing to use two different kinds of ‘P’ in one sentence).

**Exercise 9.8**

See the previous answer.

**Exercise 9.9**

Use [umsu.de/trees/](http://umsu.de/trees/).

**Exercise 9.10**

We assume that some branch on a tree contains nodes  $\eta_1 = \eta_2$  and  $A$ . We have to show that we can add  $A[\eta_1//\eta_2]$  without using the second version of Leibniz’ Law.

- k.  $\eta_1 = \eta_2$
- n.  $A$
- m.  $\eta_1 = \eta_1$  (SI)
- m+1.  $\eta_2 = \eta_1$  (k, m, LL (first version))
- m+2.  $A[\eta_1//\eta_2]$  (m+1, n, LL (first version))

**Exercise 9.11**

(a)

- 1.  $a = a$  (SI)
- 2.  $\forall x x \neq a \rightarrow a \neq a$  (UI)
- 3.  $\neg \forall x x \neq a$  (1, 2, CPL)
- 4.  $\neg \exists x x = a \leftrightarrow \forall x x \neq a$  ( $\forall \exists$ )
- 5.  $\exists x x = a$  (3, 4, CPL)
- 6.  $\Box \exists x x = a$  (5, Nec)

- (b) There are many correct answers. For example: historians debate whether Homer ever existed. If  $a$  translates ‘Homer’ then  $\exists x x = a$  is arguably false if Homer isn’t a real person. Since the available evidence is compatible with  $\neg \exists x x = a$ , the sentence  $\Box \exists x x = a$  is false on an epistemic interpretation of the box.

Where does the proof go wrong? Each of steps 1, 2, and 6 might be blamed.

**Exercise 9.12**

- (a)  $\exists x \exists y (Fx \wedge Fy \wedge x \neq y \wedge \forall z (Fz \rightarrow (z = x \vee z = y)))$   
 (b)  $\forall x \forall y \forall z \forall v (Fx \wedge Fy \wedge Fz \wedge Fv \rightarrow (x = y \vee x = z \vee x = v \vee y = z \vee y = v \vee z = v))$

**Exercise 9.13**

The *de dicto* reading of (a) can be translated as

$$\Diamond \exists x (Px \wedge \forall y (Py \rightarrow x = y) \wedge x = c),$$

where ' $P$ ' translates ' $-$  is 45th US President' and ' $c$ ' denotes Hillary Clinton. The *de re* reading can be translated as

$$\exists x (Px \wedge \forall y (Py \rightarrow x = y) \wedge \Diamond x = c).$$

The answers to (b) and (c) are analogous.

## Chapter 10

### Exercise 10.1

(a), (b), (d), and (f) are true; (c) and (e) are false.

### Exercise 10.2

Use [umsu.de/trees/](http://umsu.de/trees/). Note that the website uses slightly different identity rules: instead of the Self-Identity rule, it has a rule for closing any branch that contains a statement of the form  $\tau \neq \tau$ .

### Exercise 10.3

- (a)  $W = \{w\}$ ,  $wRw$ ,  $D = \{\text{Alice}\}$ ,  $V(F, w) = \{\text{Alice}\}$ ,  $V(G, w) = \emptyset$
- (b)  $W = \{w, v\}$ ,  $wRw$  and  $wRv$ ,  $D = \{\text{Alice}, \text{Bob}\}$ ,  $V(F, w) = \{\text{Alice}\}$ ,  $V(F, v) = \{\text{Bob}\}$
- (c)  $W = \{w, v\}$ ,  $wRw$  and  $wRv$ ,  $D = \{\text{Alice}, \text{Bob}\}$ ,  $V(F, w) = \{\text{Alice}\}$ ,  $V(F, v) = \emptyset$
- (d)  $W = \{w, v\}$ ,  $wRw$  and  $wRv$ ,  $D = \{\text{Alice}, \text{Bob}\}$ ,  $V(P, w) = \{\text{Alice}\}$ ,  $V(P, v) = \emptyset$ ,  
 $V(Q, w) = \{\text{Alice}\}$ ,  $V(Q, v) = \emptyset$

### Exercise 10.4

$\Box \forall x \exists y x = y \rightarrow \forall x \Box \exists y x = y$  is an instance of the Converse Barcan Formula. If we read the box as a relevant kind of circumstantial necessity, and Loafy could have failed to exist, then the consequent of this conditional is false. But the antecedent is true.

### Exercise 10.5

(1) is equivalent to the Barcan Formula, (3) to the Converse Barcan Formula. (2) is highly implausible. (1) and (3) are often regarded as implausible, for the reasons I discuss in the text.

Curiously, (4) seems to be equivalent to the Converse Barcan Formula: it, too, is valid on a frame iff the frame has increasing domains. It also rules out scenarios in which individuals at one world may fail to exist at an accessible world.

### Exercise 10.6

- (a)
- |    |   |     |        |
|----|---|-----|--------|
| 1. | $\exists x \Box Fx \rightarrow \Box \exists x Fx$ | (w) | (Ass.) |
| 2. | $\exists x \Box Fx$                               | (w) | (1)    |
| 3. | $\neg \Box \exists x Fx$                          | (w) | (1)    |
| 4. | $\Box Fa$   | (w) | (2)    |
| 5. | $wRv$   |     | (3)    |
| 6. | $\neg \exists x Fx$                               | (v) | (3)    |
| 7. | $Fa$  | (v) | (4,5)  |
| 8. | $a = a$   | (v) | (7)    |
- $\swarrow$   
 9.  $a \neq a$  (v) (6)  
 x

$\searrow$   
 9.  $\neg Fa$  (v) (6)  
 x

(b) DIY. The tree has four branches. I can't typeset it.

- (c)
- |    |                             |     |        |
|----|-----------------------------|-----|--------|
| 1. | $\neg \Box \exists x x = x$ | (w) | (Ass.) |
| 2. | $wRv$                       |     | (1)    |
| 3. | $\neg \exists x x = x$      | (v) | (1)    |
| 4. | $a = a$                     | (v) | (Ex.)  |
- $\swarrow$   
 9.  $a \neq a$  (v) (3)  
 x

$\searrow$   
 9.  $a \neq a$  (v) (3)  
 x

- (d)
- |    |   |     |        |
|----|---|-----|--------|
| 1. | $\neg(\Diamond Fa \rightarrow \Diamond \exists x Fx)$ | (w) | (Ass.) |
| 2. | $\Diamond Fa$   | (w) | (1)    |
| 3. | $\neg \Diamond \exists x Fx$                          | (w) | (1)    |
| 4. | $wRv$   |     | (2)    |
| 5. | $Fa$  | (v) | (2)    |
| 6. | $a = a$   | (v) | (5)    |
| 7. | $\neg \exists x Fx$                                   | (v) | (3,4)  |
- $\swarrow$   
 9.  $a \neq a$  (v) (3)  
 x

$\searrow$   
 10.  $\neg Fa$  (v) (3)  
 x

(e) 1.	$\neg(a = b \rightarrow \Box(a = a \rightarrow a = b))$	(w) (Ass.)
2.	$a = b$	(w) (1)
3.	$\neg\Box(a = a \rightarrow a = b)$	(w) (1)
4.	$wRv$	(3)
5.	$\neg(a = a \rightarrow a = b)$	(v) (3)
6.	$a = a$	(v) (5)
7.	$\neg a = b$	(v) (5)
8.	$a = b$	(v) (2,6)
	x	

### Exercise 10.7

In the definition of a model, we could allow the interpretation function to be undefined for some names. We might also allow the sets  $D_w$  to be empty. In the truth definition 10.4, we only need to clarify that  $M, w, g \not\models A$  for every atomic sentence  $A$  that contains a term  $\tau$  for which  $[\tau]^{M,g}$  is undefined.

### Exercise 10.8

In the Superman case, Clark Kent and Superman are the same person, but Lois Lane doesn't know that they are. So we appear to have  $c = s$  but not  $\Box c = s$ . Similarly, in the Julius case, Julius and Whitcomb L. Judson are the same person, but one may well not know that they are. In the Goliath case, we have  $\text{Lumpl} = \text{Goliath}$  without it being metaphysically necessary that  $\text{Lumpl} = \text{Goliath}$ , as there are worlds in which Lumpl is a bowl and Goliath is not.

### Exercise 10.9

'Lumpl' might express a concept that maps every world  $w$  to a certain piece of clay at  $w$ , where that piece is perhaps individuated by its matter or origin. The piece's shape doesn't matter. 'Goliath' might instead express a concept that maps every world  $w$  to a certain statue at  $w$ , where the statue is perhaps individuated by its shape and origin.

### Exercise 10.10

If we treat 'my body' as a name, the premises are  $\Box\exists x x = i$  and  $\neg\Box\exists x x = b$ . The conclusion is  $i \neq b$ . This argument is CK-valid and VK-valid. (It is not valid in individual

concept semantics.) It might be better to translate 'my body' as a description. This would make the argument CK-invalid and VK-invalid.

**Exercise 10.11**

Translation:  $\exists x(Tx \wedge Wx \wedge \neg KWx \wedge \neg K\neg Wx)$ , where  $T$  translates '– is a ticket' and '– will win'.

If variables are directly referential, then this sentence is true in any scenario in which I don't know which ticket will win.

**Exercise 10.12**

To render  $\forall x\forall y(x = y \rightarrow \Box x = y)$  valid, we can restrict the eligible individual concepts in a model as follows. For any (eligible) individual concepts  $f$  and  $g$  and worlds  $w$  and  $v$ , if  $wRv$  and  $f(w) = g(w)$  then  $f(v) = g(v)$ . (We do not stipulate that if  $wRv$  and  $f(v) = g(v)$  then  $f(w) = g(w)$ , which would render the necessity of distinctness valid.)