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 basic idea of possible-worlds semantics is to analyze possibility and necessity in terms of truth at possible worlds:

> 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 complete way things might have been. It may help to start with an incomplete way things might have been. For example, I might have had coffee for breakfast today. (In fact I had tea.) This is an incomplete way things might have been, because it leaves many things open: how much coffee I had, which cup I used, how fast I consumed the coffee, and so on. A possible world, by contrast, leaves nothing open; it settles every question.

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

In possible-worlds semantics, different flavours of modality correspond to different conceptions of what should count as a possible world. If we are talking about epistemic possibility and necessity, we may understand a possible world to be a complete way things *might be*, in the epistemic sense of 'might'. For circumstantial possibility and necessity, where we're interested in what is compatible or entailed by certain facts, we only consider worlds in which these facts obtain. Holding fixed the

absence of a railway line between Auckland and Sydney, for example, there is no possible world at which you can travel from Auckland to Sydney by train.

Exercise 2.1

What should count as a possible world if we're interested in the deontic sense of 'necessary' and 'possible'?

Possible-worlds semantics effectively allows us to translate modal statements about possibility and necessity into quantificational statements about possible worlds. You may feel uneasy about this. Talking about worlds other than the actual world may strike you as fanciful and unscientific. Besides, can facts about possibility and necessity really be explained in terms of facts about other worlds? Intuitively, the reason why it is impossible to travel from Auckland to Sydney by train is not that there are no "circumstantially possible worlds" in which such a trip takes place. The true reason is simply that there's no railway connection between the two cities, here in our world. No need to bring in other worlds.

These are fair concerns, and they were widely discussed in the early days of possible-worlds semantics. Fortunately, we don't have to resolve them. Our topic is the logic, not the metaphysics of modality. When we use possible-world semantics, we don't assume that the translation in terms of possible worlds somehow 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.

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

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

When we go beyond modality in the application of modal logic, the "possible worlds" will look even less like genuine worlds. In temporal logic, the role of worlds

will be played by times. In the logic of relativistic spacetime, a "world" is a spacetime point. In dynamic logic (as used in computer science), the "worlds" are often states of a computer program. In the more abstract realm of pure modal logic, "worlds" are simply points relative to which sentences are true or false.

2.2 Models

We will use possible-worlds semantics to interpret our modal language $\mathfrak{L}_M$. To this end, we first need to review what logicians mean by a "model".

Recall that a sentence is *valid* if there is no conceivable scenario in which it is false. A sentence is *logically valid* if it is valid merely in virtue of its logical form. That is:

> A sentence is logically valid iff it is true in all conceivable scenarios under all interpretations of the descriptive vocabulary.

We need to make this more precise. What is a "conceivable scenario"? What is an "interpretation of the descriptive vocabulary"? A *model*, in logic, is a simplified representation of a scenario together with an interpretation of descriptive vocabulary, containing just enough information to allow for a precise definition of validity.

Take classical propositional logic, without modal operators. Here (like in modal propositional logic), the only descriptive vocabulary are the sentence letters. What do you need to know about a scenario and an interpretation of the sentence letters in order to figure out whether an arbitrary sentence is true, in that scenario, under the given interpretation?

Answer: all you need to know is which sentence letters are true and which are false in the scenario. That's because all operators in classical propositional logic are truth-functional. Once you know that p is true and q false, you know that $p \vee q$ is true and $p \rightarrow (q \wedge p)$ is false. You don't have to know *why* p is true and q is false. Perhaps p means that it is raining and q that it is snowing, and the relevant scenario is one in which it is raining. Or perhaps p means that grass is green and q means that grass is purple and the relevant scenario is one in which grass is green. It doesn't matter.

So if we want to know which sentences of classical propositional logic are true in all scenarios under all interpretations of the descriptive vocabulary, we can equivalently ask which sentences are true under all assignments of truth-values to

the sentence letters. Any such assignment is a *model* of classical propositional logic. It is a simplified representation of a scenario together with an interpretation of the descriptive vocabulary.

You may wonder why we need to consider alternative scenarios *and* alternative interpretations of the descriptive vocabulary. Why don't we say that a sentence is valid iff it is true under every interpretation of the descriptive vocabulary, without also looking at alternative scenarios?

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

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

Back to modal logic. What do we need to know about a scenario and an interpretation of the sentence letters to figure out whether an arbitrary $\mathfrak{L}_M$-sentence is true or false? It is usually not enough to know the truth-values of the sentence letters, because the box and the diamond are (usually) not understood to be truth-functional. Knowing that p is true doesn't tell you whether $\Box p$ is true. We need more information about the relevant scenario and about the meaning of p.

You might suggest that we need to know not only which sentence letters are true, but also which of them are necessarily (or possibly) true. But that still wouldn't be enough: if you know that neither p nor q is necessary, you generally can't determine whether, say, $\Box(p \to q)$ is true.

We could take a model to explicitly classify even complex sentences as necessary or possible. But that would be too liberal. It would allow for models in which $\Box(p \wedge q)$ is true while $\Box p$ is false. In most applications of modal logic, we want $\Box(p \wedge q) \to \Box p$ to be valid.

This is where possible-worlds semantics will help us. We will assume that any relevant scenario consists of a range of possible worlds. Sentence letters are true or false relative to these worlds. Within each world, the rules of classical logic hold.

For example, if p and q are both true at a world, then so is $p \wedge q$. So we don't have to explicitly specify at which worlds logically complex sentences are true. This carries over to modal sentences: $\Box p$, for example, is true iff p is true at all worlds.

Formally, we therefore define a model of modal propositional logic as consisting of two parts. First, there is a set of things we call "worlds". Second, there is an interpretation function that assigns a truth-value to each sentence letter at each world.

Definition 2.1

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

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

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

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

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

As announced above, V only explicitly fixes the truth-values of sentence letters. The truth-value of complex sentences at a world is determined by the meaning of the logical operators.

To specify how the truth-value of complex sentences are determined, I need another piece of notation. 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 '$M, w \not\models A$' as the negation of '$M, w \models A$'.

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

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

Definition 2.2: Basic Possible-Worlds Semantics

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

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

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

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

Clauses (c)–(f) similarly tell us that the truth-tables for the other truth-functional connectives apply locally at each world.

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

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

Every $\mathfrak{L}_M$-sentence is built up from sentence letters with the operators covered in definition 2.2. Hence the definition settles the truth-value of every sentence at every world in every model.

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

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

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

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

Exercise 2.2

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

2.3 Validity

We can now replace the hand-wavy definition of validity in terms of "conceivable scenarios" by a precise definition in terms of models: a sentence is valid iff it is true in all models.

However, there's a small problem. The way we've defined models, $\mathfrak{L}_M$-sentences are generally true or false only relative to a model *and a world*. Intuitively, the scenarios represented by our models often contain many possible worlds. And if a scenario contains more than one world, the truth-value of a sentence will often vary from world to world. We could fix this by changing the definition of a model so that a model must designate a particular member of W as representing the actual world. Formally, a model would be a triple of a set W, an interpretation V, and a member w of W. Such models are called *pointed models*. With pointed models, we could say that a sentence is true in a model iff it is true at the actual world of the model.

An alternative (but ultimately equivalent) approach, which we will follow, is to stipulate that a sentence is **true in a model** (as opposed to at a world in a model) iff it is true at all worlds in the model. Validity therefore amounts to truth at every world in ever model.

Definition 2.3

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

Validity is a special case of logical consequence. The general case is defined as follows.

Definition 2.4

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

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

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

$$A \wedge B \models A$$

expresses that every instance of the schema $A \wedge B$ logically entails the corresponding

instance of A. Similarly, the schematic

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

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

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

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

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

Propositional Extension Theorem

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

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

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

Why is that?

> *Very rough proof sketch.* By definition 2.4, '$p, p \rightarrow q \models q$' means that whenever p and $p \rightarrow q$ are true at a world in a model, then so is q. Now look at definition 2.2. According to clause (e), $p \rightarrow q$ is true at a world iff p is false at that world or q is true. So if p and $p \rightarrow q$ are both true at a world, then q can't be false at that world.
>
> In general, the Propositional Extension Theorem holds because the truth-functional connectives have their standard truth-table meaning relative to every world. If the standard truth-tables for the connectives guarantee that A is true

whenever all members of Γ are true, it therefore follows that A is true at any world at which all members of Γ are true. □

Lastly, on p.15, I made the following claim:

Replacement Theorem

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

Proof sketch. Given definition 2.3, the Replacement Theorem immediately follows from the following hypothesis, which I'll call '(*)':

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

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

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

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

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

I won't bore you by going through all the remaining cases: that C is disjunction $D \vee E$, a conditional $D \rightarrow E$, a biconditional $D \leftrightarrow E$, a box sentence $\Box D$, and a

> diamond sentence $\Diamond D$. In each case, definition 2.2 entails that if (*) holds for the immediate parts of C, then (*) also holds for C. It then follows that (*) holds for all sentences whatsoever, because every $\mathfrak{L}_M$-sentence is built up from sentence letters by the operators $\neg$, $\wedge$, $\vee$, etc. □

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

2.4 The logic of unrestricted modality

By definition 2.3, a sentence is valid iff it is true at all worlds in all models. Definition 2.1 tells us what counts as a model; definition 2.2 determines the truth-value of any sentence at any world in any model. Together, these definitions therefore settle which sentences, and which schemas, are valid and which aren't.

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

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

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

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

(**4**) $\quad \Box A \rightarrow \Box\Box A$

If something is necessary, is it necessarily necessary? Our possible-worlds semantics

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

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

Exercise 2.3

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

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

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

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

Exercise 2.4

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

Make sure you don't conflate the concepts of necessity and validity. Necessity means truth at all worlds (or so we currently assume). Validity means truth at all worlds *in all models*. Whether an $\mathfrak{L}_M$ sentence is necessary generally varies from

model to model. In a model whose interpretation function assigns 1 to p relative to each world, p is necessary insofar as $\Box p$ is true at all worlds in the model. In other models, $\Box p$ is not true at all worlds. Validity, by contrast, is not relative to a model. The sentence p is definitely not valid. The sentence $\Box p \rightarrow p$ is.

Exercise 2.5

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

Here is an example of an invalid schema:

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

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

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

How could we show that this isn't valid?

By definition 2.3, 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 above sentence is false. Such a model is called a **countermodel** for the sentence (or schema) we want to reveal as invalid.

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

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

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

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

Exercise 2.6

First show that the schema $A \rightarrow \Box A$ is invalid. So we do not have $A \models \Box A$. Explain why this doesn't contradict the previous exercise.

Exercise 2.7

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

A brief aside: In the previous chapter, we saw that different applications of modal logic call for different standards of validity. If the box formalizes knowledge, $\Box A \rightarrow A$ should count as valid, but not if the box formalizes obligation. The semantics of the present chapter only gives us one kind of validity – a kind that is not suitable for deontic logic. That's why we'll generalize the semantics in the next chapter.

2.5 Trees

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

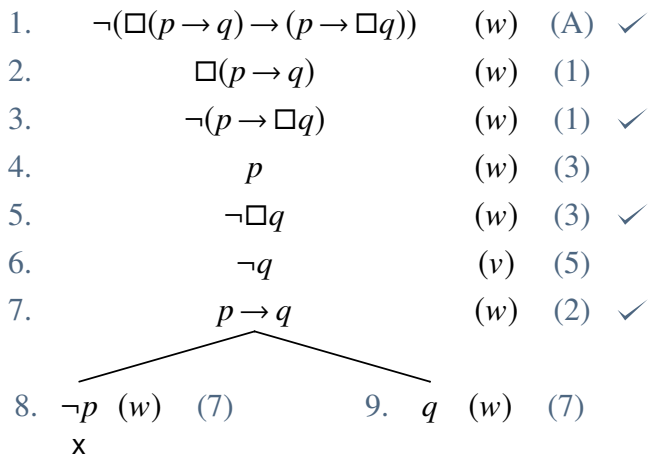
6.	$\neg q$	(v)	(5)

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

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

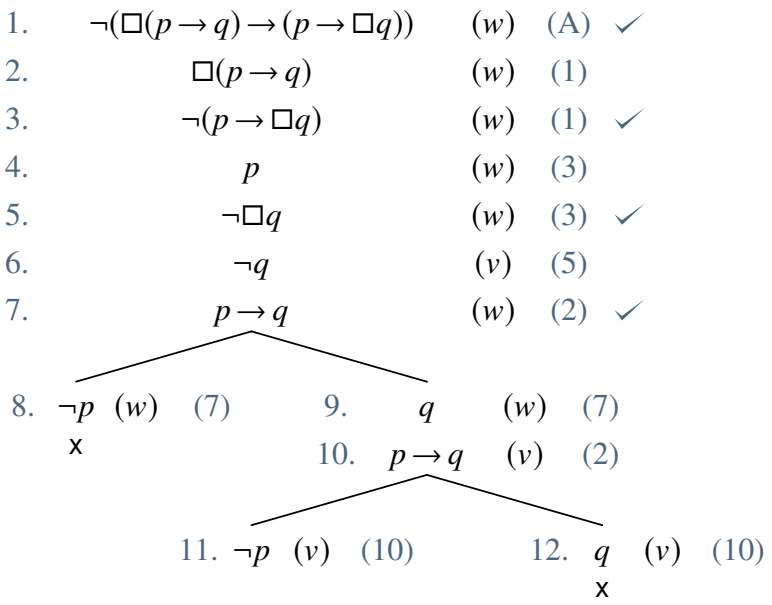
If you are used to "natural deduction" proofs, you may now be tempted to apply *Modus Ponens* and infer that q is true at w, from lines 4 and 7. In the tree method, however, we try not to draw inferences from multiple premises. We simply look at any node that has not yet been processed and check what it tells us about the sub-sentences it contains. So we process node 7 without looking at node 4.

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



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

The countermodel is given by all the assumptions on the open middle branch. Here we find two worlds, $W = \{w, v\}$. The interpretation function V makes p true at w (node 4) and false at v (node 10); q is false at both v and w (nodes 6 and 9).

Now for the general rules.

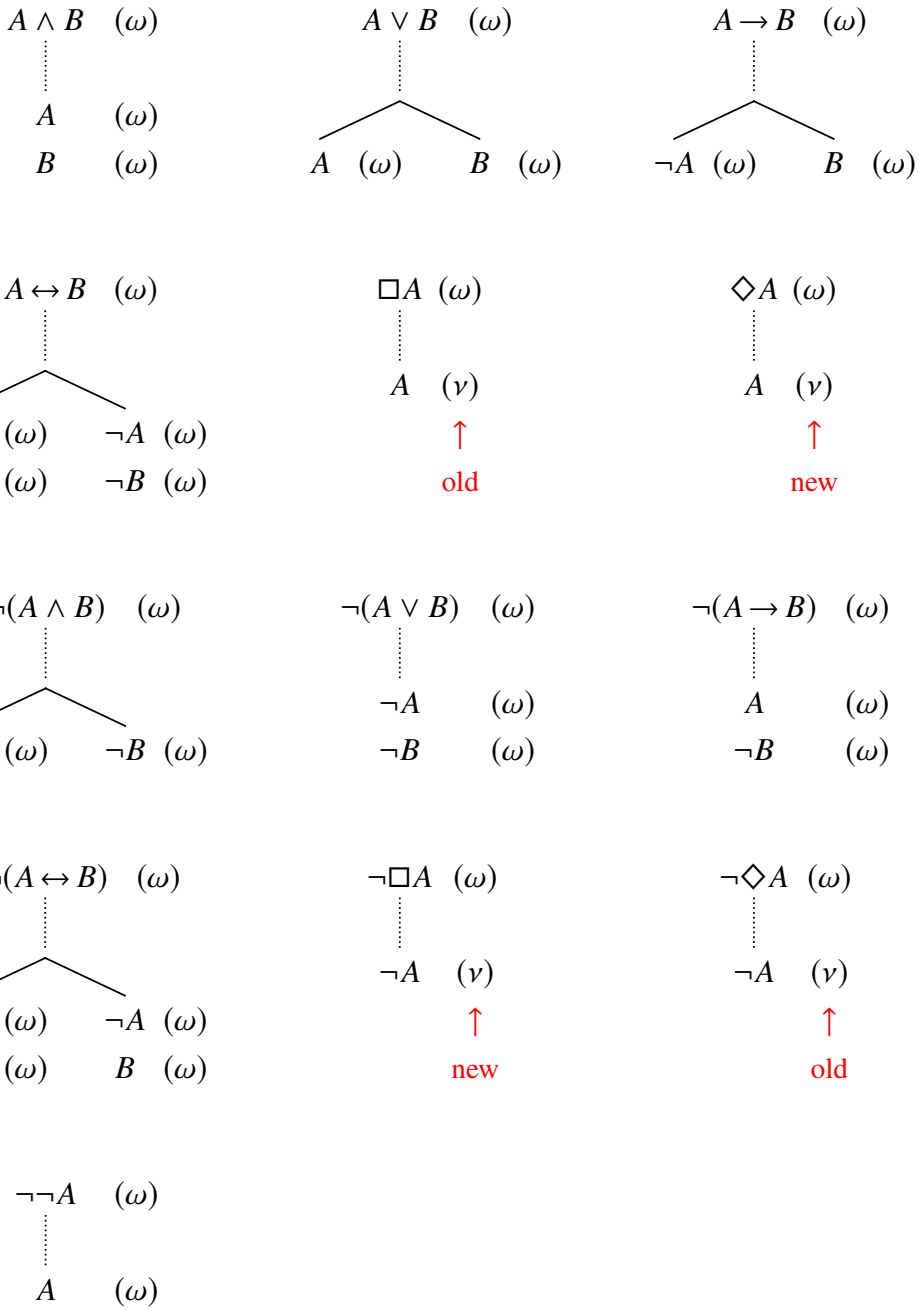
In order to find a countermodel for a sentence A with the help of the tree method, you always begin by assuming that the *negation* of A is true at world w:

$$1. \quad \neg A \quad (w) \quad (\text{A})$$

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

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

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



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

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

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

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

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

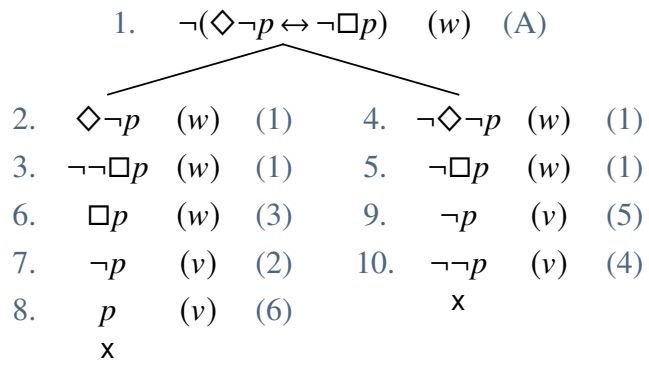
Exercise 2.8

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

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

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

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



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

Exercise 2.9

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

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

Exercise 2.10

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

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

Exercise 2.11

Can we use the tree method to show that premises $A_1, \ldots, A_n$ logically entail conclusion B? If so, how?