Preface

This book is an introduction to logic for students of contemporary philosophy. It covers (i) basic approaches to logic, including proof theory and especially model theory, (ii) extensions of standard logic (such as modal logic) that are important in philosophy, and (iii) some elementary philosophy of logic. It prepares students to read the logically sophisticated articles in today's philosophy journals, and helps them resist bullying by symbol-mongers. In short, it teaches the logic you need to know in order to be a contemporary philosopher.

For better or for worse (I think better), during the last century or so, philosophy has become infused with logic. Logic informs nearly every area of philosophy; it is part of our shared language and knowledge base. The standard philosophy curriculum therefore includes a healthy dose of logic. This is a good thing. But in many cases only a single advanced logic course is required; and the material taught in that course becomes the only advanced logic that many undergraduate philosophy majors and beginning graduate students ever learn. And this one course is often an intensive survey of metalogic (for example, one based on the excellent Boolos et al. (2007)). I do believe in the value of such a course, especially for students who take multiple logic courses or specialize in "technical" areas of philosophy. But for students taking only a single course, that course should not, I think, be a course in metalogic. The standard metalogic course is too mathematically demanding for the average philosophy student, and omits material that the average student ought to know. If there can be only one, let it be a crash course in logic literacy.

"Logic literacy" includes knowing what metalogic is all about. And you can't really learn about anything in logic without getting your hands dirty and doing it. So this book does contain some metalogic (for instance, soundness and completeness proofs in propositional logic and propositional modal logic). But it doesn't cover the central metalogical results one normally covers in a mathematical logic course: soundness and completeness in predicate logic, computability, Gödel's incompleteness theorems, and so on.

Chapter 1

What is Logic?

Since you are reading this book, you probably know some logic already. You probably know how to translate English sentences into symbolic notation, into propositional logic:

English	*Propositional logic*
Either violets are blue or I need glasses	$V \vee N$
If snow is white then grass is not green	$S \rightarrow \sim G$

and into predicate logic:

English	*Predicate logic*
If Grant is male then someone is male	$Mg \rightarrow \exists x Mx$
Any friend of Barry is either insane or friends with everyone	$\forall x[Fxb \rightarrow (Ix \vee \forall y Fxy)]$

You are probably also familiar with some techniques for evaluating arguments written out in symbolic notation. You have probably encountered truth tables, and some form of proof theory (perhaps a "natural deduction" system; perhaps "truth trees".) You may have even encountered some elementary model theory. In short: you have taken an introductory course in symbolic logic.

What you already possess is: literacy in elementary logic. What you will get out of this book is: literacy in the rest of logic that philosophers tend to presuppose, plus a deeper grasp of what logic is all about.

So what *is* logic all about?

1.1 Logical consequence and logical truth

Logic is about many things, but most centrally it is about *logical consequence*. The statement "someone is male" is a logical consequence of the statement "Grant is male". If Grant is male, then it *logically follows* that someone is male. Put another way: the statement "Grant is male" *logically implies* the statement "someone is male". Likewise, the statement "Grant is male" is a logical consequence of the statements "It's not the case that Leisel is male" and "Either Leisel is male or Grant is male" (taken together). The first statement *follows from* the latter two statements; they logically imply it. Put another way: the argument whose premises are the latter two statements, and whose conclusion is the former statement, is a *logically correct* one.[1]

So far we've just given synonyms. The following slogan advances us a bit further: logical consequence is *truth-preservation by virtue of form*. To unpack a bit: for ϕ to be a logical consequence of ψ, it is not enough that we all know that ϕ is true if ψ is. We all know that an apple will fall if it is dropped, but the relationship between falling and dropping does not hold by virtue of logic. Why not? For one thing, "by virtue of logic" requires the presence of some sort of necessary connection, a connection that is absent in the case of the dropped apple (since it would be possible—in some sense—for a dropped apple not to fall). For another, it requires the relationship to hold by virtue of the forms of the statements involved, whereas the relationship between "the apple was dropped" and "the apple fell" holds by virtue of the contents of these statements and not their form. (By contrast, the inference from 'It's not the case that Leisel is male" and "Either Leisel is male or Grant is male" to "Grant is male" is said to hold in virtue of form, since any argument of the form "it's not the case that ϕ; either ϕ or ψ; therefore ψ" is logically correct.) As we'll see shortly, there are many open philosophical questions in this vicinity, but perhaps we have enough of an intuitive fix on the concept of logical consequence to go on with, at least for the moment.

A related concept is that of a *logical truth*. Just as logical consequence is truth-preservation by virtue of form, logical truth is *truth by virtue of form*. Examples might include: "it's not the case that snow is white and also not white", "All fish are fish", and "If Grant is male then someone is male". As with logical consequence, logical truth is thought to require some sort of necessity

[1] The word 'valid' is sometimes used for logically correct arguments, but I will reserve that word for a different concept: that of a logical truth, under the semantic conception.

and to hold by virtue of form, not content. It is plausible that logical truth and logical consequence are related thus: a logical truth is a sentence that is a logical consequence of the empty set of premises. One can infer a logical truth by using logic alone, without the help of any premises.

A central goal of logic, then, is to study logical truth and logical consequence. But the contemporary method for doing so is somewhat indirect. As we will see in the next section, instead of formulating claims about logical consequence and logical truth themselves, modern logicians develop formal models of how those concepts behave.

1.2 Formalization

Modern logic is called "mathematical" or "symbolic" logic, because its method is the mathematical study of formal languages. Modern logicians use the tools of mathematics (especially, the tools of very abstract mathematics, such as set theory) to treat sentences and other parts of language as mathematical objects. They define up formal languages, define up sentences of the languages, define up properties of the sentences, and study those properties. Mathematical logic was originally developed to study mathematical reasoning, but its techniques are now applied to reasoning of all kinds.

Take propositional logic, the topic of chapter 2. Here our goal is to shed light on the logical behavior of 'and', 'or', and so on. But rather than studying those words directly, we will develop a certain formal language, the language of propositional logic. The sentences of this language look like this:

$$P$$

$$(Q \rightarrow R) \vee (Q \rightarrow \sim S)$$

$$P \leftrightarrow (P \wedge Q)$$

Symbols like $\wedge$ and $\vee$ represent natural language logical words like 'and' and 'or'; and the sentence letters $P, Q, \dots$ represent declarative natural language sentences. We will then go on to define (as always, in a mathematically rigorous way) various concepts that apply to the sentences in this formal language. We will define the notion of a *tautology* ("all Trues in the truth table"), for example, and the notion of a *provable formula* (we will do this using a system of deduction with rules of inference; but one could use truth trees, or some other method). These defined concepts are "formalized versions" of the concepts of logical consequence and logical truth.

Formalized logical consequence and logical truth should be distinguished from the real things. The formal sentence $P{\rightarrow}P$ is a tautology, but since it is uninterpreted, we probably shouldn't call it a logical truth. Rather, it *represents* logical truths like "If snow is white then snow is white". A logical truth ought at least to be *true*, after all, and $P{\rightarrow}P$ isn't true, since it doesn't even have a meaning—what's the meaning of P? (Caveat: one might *give* meanings to formal sentences—by translation into natural language ("let P mean that snow is white; let $\wedge$ mean *and*…"), or perhaps by some direct method if no natural language translation is available. And we may indeed speak of logical truth and logical consequence for *interpreted* formal sentences.)

Why are formal languages called "formal"? (They're also sometimes called "artificial" languages.) Because their properties are mathematically stipulated, rather than being pre-existent in flesh-and-blood linguistic populations. We stipulatively define a formal language's grammar. (Natural languages like English also have grammars, which can be studied using mathematical techniques. But these grammars are much more complicated, and are discovered rather than stipulated.) And we must stipulatively define any properties of the symbolic sentences that we want to study, for example, the property of being a tautology. (Sentences of natural languages already have meanings, truth values, and so on; we don't get to stipulate these.) Further, formal languages often contain abstractions, like the sentence letters $P, Q, \ldots$ of propositional logic. A given formal language is designed to represent the logical behavior of a select few natural language words; when we use it we abstract away from all other features of natural language sentences. Propositional logic, for example, represents the logical behavior of 'and', 'or', and a few other words. When a sentence contains none of these words of interest, we represent it with one of the sentence letters $P, Q, \ldots$, indicating that we are ignoring its internal structure.

1.3 Metalogic

There are many reasons to formalize—to clarify meaning, to speak more concisely, and so on. But one of the most powerful reasons is to do *metalogic*.

In introductory logic one learns to *use* certain logical systems—how to construct truth tables, derivations and truth trees, and the rest. But logicians do not develop systems only to sit around all day using them. As soon as a logician develops a new system, she begins to ask questions *about* that system. For an analogy, imagine people who make up new games for a living. If they invent a new version of chess, they might spend some time actually playing it. But if they are like logicians, they will quickly tire of this and start asking questions *about* the game. “Is the average length of this new game longer than the average length of a game of standard chess?”. “Is there any strategy that guarantees victory?” Analogously, logicians ask questions about logical systems. “What formulas can be proven in such and such a system?” “Can you prove the same things in this system as in system X?” “Can a computer program be written to determine whether a given formula is provable in this system?” The study of such questions *about* formal systems is called “metalogic”.

The best way to definitively answer metalogical questions is to use the methods of mathematics. And to use the methods of mathematics, we need to have rigorous definitions of the crucial terms that are in play. For example, in chapter 2 we will mathematically demonstrate that “every formula that is provable (in a certain formal system) is a tautology”. But doing so requires carefully defining the crucial terms: ‘formula’, ‘provable’, and ‘tautology’; and the best way to do this is to formalize. We treat the languages of logic as mathematical objects so that we can mathematically demonstrate facts about them.

Metalogic is a fascinating and complex subject; and other things being equal, it’s good to know as much about it as you can. Now, other things are rarely equal; and the premise of this book is that if push sadly comes to shove, limited classroom time should be devoted to achieving logic literacy rather than a full study of metalogic in all its glory. But still, logic literacy *does* require understanding metalogic: understanding what it is, what it accomplishes, and how one goes about doing it. So we will be doing a decent amount of metalogic in this book. But not too much, and not the harder bits.

Much of metalogic consists of proving things about formal systems. And sometimes, those formal systems themselves concern proof. For example, as I said a moment ago, we will prove in chapter 2 that every provable formula is a tautology. If this seems dizzying, keep in mind that ‘proof’ here is being used in two different senses. There are *metalogic proofs*, and there are *proofs in formal systems*. Metalogic proofs are phrased in natural language (perhaps augmented with mathematical vocabulary), and employ informal (though rigorous!) reasoning of the sort one would encounter in a mathematics book. The chapter 2 argument that “every provable formula is a tautology” will be a metalogic proof. Proofs in formal systems, on the other hand, are phrased using sentences of formal languages, and proceed according to prescribed formal rules. ‘Provable’ in the statement ‘every provable formula is a tautology’ signifies proof in a certain formal system (one that we will introduce in chapter 2), not metalogic proof.

1.8 Set theory

I said earlier that modern logic uses "mathematical techniques" to study formal languages. The mathematical techniques in question are those of set theory. Only the most elementary set-theoretic concepts and assumptions will be needed, and you may already be familiar with them; but nevertheless, here is a brief overview.

Sets have *members*. Consider the set, A, of even integers between 2 and 6. 2 is a member of A, 4 is a member of A, 6 is a member of A; and nothing else is a member of A. We use the expression "$\in$" for membership; thus, we can say: $2 \in A$, $4 \in A$, and $6 \in A$. We often name a set by putting names of its members between braces: "$\{2,4,6\}$" is another name of A.

We can also speak of sets with infinitely many members. Consider $\mathbb{N}$, the set of natural numbers. Each natural number is a member of $\mathbb{N}$; thus, $0 \in \mathbb{N}, 1 \in \mathbb{N}$, and so on. We can informally name this set with the brace notation as well: "$\{0,1,2,3,\dots\}$", so long as it is clear which continued series the ellipsis signifies.

The members of a set need not be mathematical entities; anything can be a member of a set.[8] Sets can contain people, or cities, or—to draw nearer to our intended purpose—sentences and other linguistic entities.

There is also the empty set, $\varnothing$. This is the one set with no members. That is, for each object u, u is not a member of $\varnothing$ (i.e.: for each $u, u \notin \varnothing$.)

Though the notion of a set is an intuitive one, the Russell Paradox (discovered by Bertrand Russell) shows that it must be employed with care. Let R be the set of all and only those sets that are not members of themselves. That is, R is the set of non-self-members. Russell asks the following question: is R a member of itself? There are two possibilities:

- $R \notin R$. Thus, R is a non-self-member. But R was said to be the set of all non-self-members, and so we'd have $R \in R$. *Contradiction*.

[8] Well, some axiomatic set theories bar certain "very large collections" from being members of sets. This issue won't be relevant here.

- $R \in R$. So R is *not* a non-self-member. R, by definition, contains *only* non-self-members. So $R \notin R$. *Contradiction.*

Thus, each possibility leads to a contradiction. But there are no remaining possibilities—either R is a member of itself or it isn't! So it looks like the very idea of sets is paradoxical.

Since Russell's time, set theorists have developed theories of sets that avoid Russell's paradox (as well as other related paradoxes). They do this chiefly by imposing rigid restrictions on when sets exist. So far we have been blithely assuming that there exist various sets: the set $\mathbb{N}$, sets containing people, cities, and sentences, Russell's set R. That got us into trouble. So what we want is a theory of when sets exist that blocks the Russell paradox by saying that set R simply doesn't exist (for then Russell's argument falls apart), but which says that the sets we need to do mathematics and metalogic *do* exist. The details of set theory are beyond the scope of this book. Here, we will help ourselves to intuitively "safe" sets, sets that aren't anything like the Russell set. We'll leave the task of what "safe" amounts to, exactly, to the set theorists.

Various other useful set-theoretic notions can be defined in terms of the notion of membership. Set A is a *subset* of set B ("$A \subseteq B$") when every member of A is a member of B. The *intersection* of A and B ("$A \cap B$") is the set that contains all and only those things that are members of both A and B; the *union* of A and B ("$A \cup B$") is the set containing all and only those things that are members of either A or B (or both[9]).

Suppose we want to refer to the set of the so-and-sos—that is, the set containing all and only objects, u, that satisfy the condition "so-and-so". We'll do this with the term "$\{u$: u is a so-and-so$\}$". Thus, we could write: "$\mathbb{N} = \{u :$ u is a natural number$\}$". And we could restate the definitions of $\cap$ and $\cup$ from the previous paragraph as follows:

$$A \cap B = \{u : u \in A \text{ and } u \in B\}$$
$$A \cup B = \{u : u \in A \text{ or } u \in B\}$$

Sets have members, but they don't contain them in any particular order. For example, the set containing me and Barack Obama doesn't have a "first" member. "{Ted, Obama}" and "{Obama, Ted}" are two different names for the same set—the set containing just Obama and me. (This follows from

[9] In this book I always use 'or' in its inclusive sense.

the "criterion of identity" for sets: sets are identical if and only if they have exactly the same members.) But sometimes we need to talk about set-*like* things containing objects in a particular order. For this purpose we use *ordered sets*.[10] Two-membered ordered sets are called ordered pairs. To name the ordered pair of Obama and Ted, we use: "⟨Obama, Ted⟩". Here, the order is significant; ⟨Obama, Ted⟩ and ⟨Ted, Obama⟩ are *not* the same ordered pair. The three-membered ordered set of $u, v,$ and w (in that order) is written: $\langle u, v, w\rangle$; and similarly for ordered sets of any finite size. A n-membered ordered set is called an *n-tuple*. (For the sake of convenience, let's define the 1-tuple $\langle u\rangle$ to be just the object u itself.)

A further concept we'll need is that of a *relation*. A relation is just a feature of multiple objects taken together. The taller-than relation is one example: when one person is taller than another, that's a feature of those two objects taken together. Another example is the less-than relation for numbers. When one number is less than another, that's a feature of those two numbers taken together.

"Binary" relations apply to two objects at a time. The taller-than and less-than relations are binary relations, or "two-place" relations as we might say. We can also speak of three-place relations, four-place relations, and so on. An example of a three-place relation would be the *betweenness* relation for numbers: the relation that holds among 2, 5, and 23 (in that order), for example.

We can use ordered sets to give an official definition of what a relation is.

Definition of relation: An n-place relation is a set of n-tuples.

So a binary (two-place) relation is a set of ordered pairs. For example, the taller-than relation may be taken to be the set of ordered pairs $\langle u, v\rangle$ such that u is a taller person than v. The *less-than* relation for positive integers is the set of ordered pairs $\langle m, n\rangle$ such that m is a positive integer less than n, another positive integer. That is, it is the following set:

$$\{\langle 1,2\rangle, \langle 1,3\rangle, \langle 1,4\rangle \dots \langle 2,3\rangle, \langle 2,4\rangle \dots\}$$

[10] There's a trick for defining ordered sets in terms of sets. First, define the ordered pair $\langle u,v\rangle$ as the set $\{\{u\},\{u,v\}\}$. (We can recover the information that u is intended to be the *first* member because u "appears twice".) Then define the n-tuple $\langle u_1 \dots u_n\rangle$ as the ordered pair $\langle u_1, \langle u_2 \dots u_n\rangle\rangle$, for each $n \geq 3$. But henceforth I'll ignore this trick and just speak of ordered sets without worrying about how they're defined.

When $\langle u, v\rangle$ is a member of relation R, we say, equivalently, that u and v "stand in" R, or R "holds between" u and v, or that u "bears" R to v. Most simply, we write "Ruv".[11]

Some more definitions:

Definition of domain, range, over: Let R be any binary relation and A be any set.

- The domain of R ("dom(R)") is the set $\{u\colon \text{for some } v, Ruv\}$
- The range of R ("ran(R)") is the set $\{u\colon \text{for some } v, Rvu\}$
- R is over A iff $\text{dom}(R) \subseteq A$ and $\text{ran}(R) \subseteq A$

In other words, the domain of R is the set of all things that bear R to something; the range is the set of all things that something bears R to; and R is over A iff the members of the 'tuples in R are all drawn from A.

Binary relations come in different kinds, depending on the patterns in which they hold:

Definition of kinds of binary relations: Let R be any binary relation over some set A.

- R is serial (in A) iff for every $u \in A$, there is some $v \in A$ such that Ruv.
- R is reflexive (in A) iff for every $u \in A, Ruu$
- R is symmetric iff for all u, v, if Ruv then Rvu
- R is transitive iff for any u, v, w, if Ruv and Rvw then Ruw
- R is an equivalence relation (in A) iff R is symmetric, transitive, and reflexive (in A)
- R is total (in A) iff for every $u, v \in A, Ruv$

Notice that we relativize some of these relation types to a given set A. We do this in the case of reflexivity, for example, because the alternative would be to say that a relation is reflexive *simpliciter* if *everything* bears R to itself; but that would require the domain and range of any reflexive relation to be the set of absolutely all objects. It's better to introduce the notion of being reflexive relative to a set, which is applicable to relations with smaller domains. (I will

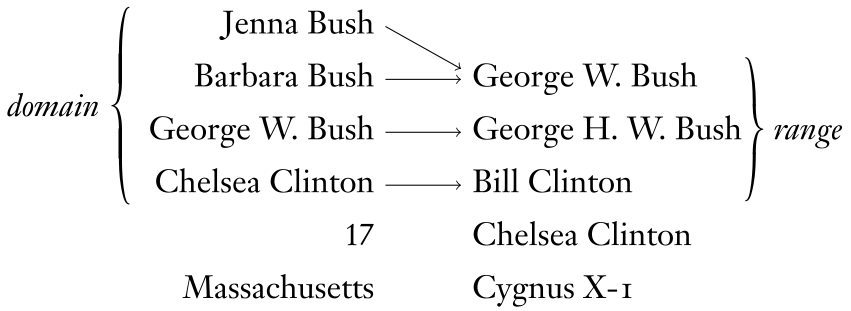
[11] This notation is like that of predicate logic; but here I'm speaking the metalanguage, not displaying sentences of a formalized language.

sometimes omit the qualifier ‘in A’ when it is clear which set that is.) Why don’t symmetry and transitivity have to be relativized to a set?—because they only say what must happen *if* R holds among certain things. Symmetry, for example, says merely that *if* R holds between u and v, then it must also hold between v and u, and so we can say that a relation is symmetric absolutely, without implying that everything is in its domain.

We’ll also need the concept of a *function*. A function “takes in” an object or objects (in a certain order), and “spits out” a further object. For example, the addition function takes in two numbers, and spits out their sum. As with sets, ordered sets, and relations, functions are not limited to mathematical entities: they can take in and spit out any objects whatsoever. We can speak of the *father-of* function, for example, which takes in a person, and spits out the father of that person. (The more common way of putting this is: the function “maps” the person to his or her father.) And later in this book we will be considering functions that take in and spit out linguistic entities.

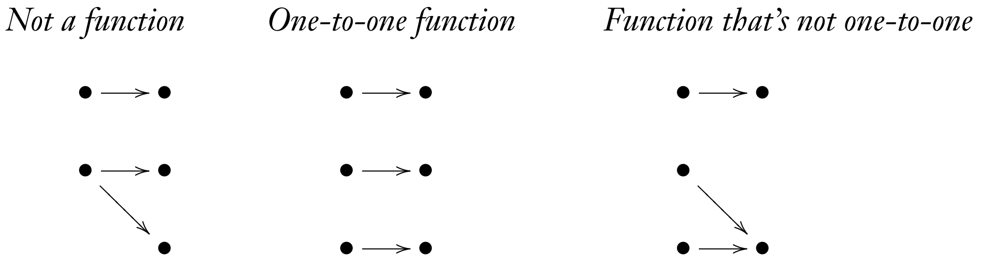
Some functions must take in more than one object before they are ready to spit out something. For example, you need to give the addition function two numbers in order to get it to spit out something; for this reason it is called a *two-place* function. The father-of function, on the other hand, needs to be given only one object, so it is a one-place function. Let’s simplify this by thinking of an n-place function as simply being a one-place function that takes in only n-tuples. Thus, if you give the addition function the ordered pair $\langle 2,5\rangle$, it spits out 7.

The objects that a function takes in are called its *arguments*, and the objects it spits out are called its *values*. If u is an argument of f we write “$f(u)$” for the value of function f as applied to the argument u. $f(u)$ is the object that f spits out, if you feed it u. For example, where f is the father-of function, since Ron is my father we can write: $f(\text{Ted}) = \text{Ron}$. When f is an n-place function—i.e., its arguments are n-tuples—instead of writing $f(\langle u_1,\dots,u_n\rangle)$ we write simply $f(u_1,\dots,u_n)$. So where a is the addition function, we can write: $a(2,3) = 5$. The *domain* of a function is the set of its arguments, and its *range* is the set of its values. If u is not in function f’s domain (i.e., u is not one of f’s arguments), then f is *undefined* for u. The father-of function, for example, is undefined for numbers (since numbers have no fathers). These concepts may be pictured for (a part of) the father-of function thus:



The number 17 and the state of Massachusetts are excluded from the domain because, being a number and a political entity, they don't have fathers. Chelsea Clinton and Cygnus X-1 are excluded from the range because, being a woman and a black hole, they aren't fathers of anyone. 17 and Massachusetts aren't in the range either; and Cygnus X-1 isn't in the domain. But Chelsea Clinton is in the domain, since she has a father.

It's part of the definition of a function that a function can never map an argument to two distinct values. That is, $f(u)$ cannot be equal both to v and also to v' when v and v' are two different objects. That is, a function always has a unique value, given any argument for which the function is defined. (So there is no such function as the *parent-of* function; people typically have more than one parent.) Functions *are* allowed to map two distinct arguments to the same value. (The father-of function is an example; two people can have the same father.) But if a given function happens never to do this, then it is called *one-to-one*. That is, a (one-place) function f is one-to-one iff for any u and v in its domain, if $u \neq v$ then $f(u) \neq f(v)$. (The function of natural numbers f defined by the equation $f(n) = n + 1$ is an example.) This all may be pictured as follows:


Not a function *One-to-one function* *Function that's not one-to-one*

As with the notion of a relation, we can use ordered sets to give official definitions of function and related notions:

Definition of function-theoretic notions:

- A function is a set of ordered pairs, f, obeying the condition that if $\langle u,v\rangle$ and $\langle u,w\rangle$ are both members of f, then $v=w$
- When $\langle u,v\rangle \in f$, we say that u is an argument of f, v is a value of f, and that f maps u to v; and we write: "$f(u)=v$"
- The domain of a function is the set of its arguments; its range is the set of its values
- A function is n-place when every member of its domain is an n-tuple

Thus, a function is just a certain kind of binary relation—one that never relates a single thing u to two distinct objects v and w. (Notice that the definition of "domain" and "range" for functions yields the same results as the definition given earlier for relations.)

The topic of infinity is perhaps set theory's most fascinating part. And one of the most fascinating things about infinity is the matter of sizes of infinity. Compare the set $\mathbb{N}$ of natural numbers and the set $\mathbb{E}$ of even natural numbers ($\{0,2,4,6,\dots\}$). Which set is bigger—which has more members? You might think that $\mathbb{N}$ has got to be bigger, since it contains all the members of $\mathbb{E}$ and then the odd natural numbers in addition. But in fact these sets have the same size. For we can line up their members as follows:

$$\begin{array}{lccccccc} \mathbb{N}: & 0 & 1 & 2 & 3 & 4 & 5 & \dots \\ \mathbb{E}: & 0 & 2 & 4 & 6 & 8 & 10 & \dots \end{array}$$

If two sets can be "lined up" in this way, then they have the same size. Indeed, this is how set theorists define 'same size'. Or rather, they give a precise definition of sameness of size (they call it "equinumerosity", or sameness of "cardinality") which captures this intuitive idea:

Definition of Equinumerosity: Sets A and B are equinumerous iff there exists some one-to-one function whose domain is A and whose range is B

Intuitively: sets are equinumerous when each member of either set can be associated with a unique member of the other set. You can line their members up.

The picture in which the members of $\mathbb{N}$ and the members of $\mathbb{E}$ were lined up is actually a picture of a function: the function that maps each member of $\mathbb{N}$

to the member of $\mathbb{E}$ immediately below it in the picture. Mathematically, this function, f, may be defined thus:

$$f(n) = 2n \qquad\qquad \text{(for any } n \in \mathbb{N}\text{)}$$

This function is one-to-one (since if two natural numbers are distinct then doubling each results in two distinct numbers). So $\mathbb{N}$ and $\mathbb{E}$ are equinumerous. It's quite surprising that a set can be equinumerous with a mere subset of itself. But that's how it goes with infinity.

Even more surprising is the fact that the rational numbers are equinumerous with the natural numbers. A (nonnegative) rational number is a number that can be written as a fraction $\frac{n}{m}$ where n and m are natural numbers and $m \neq 0$. To show that $\mathbb{N}$ is equinumerous with the set $\mathbb{Q}$ of rational numbers, we must find a one-to-one function whose domain is $\mathbb{N}$ and whose range is $\mathbb{Q}$. At first this seems impossible, since the rationals are "dense" (between every two fractions there is another fraction) whereas the naturals are not. But we must simply be clever in our search for an appropriate one-to-one function.

Each rational number is represented in the following grid:

$$
\begin{array}{cc}
 & \text{denominators} \\
\text{numerators} &
\begin{array}{c|cccccc}
 & 1 & 2 & 3 & 4 & 5 & \cdots \\
\hline
0 & \frac{0}{1} & \frac{0}{2} & \frac{0}{3} & \frac{0}{4} & \frac{0}{5} & \cdots \\
1 & \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\
2 & \frac{2}{1} & \frac{2}{2} & \boxed{\frac{2}{3}} & \frac{2}{4} & \frac{2}{5} & \cdots \\
3 & \frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \cdots \\
4 & \frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
\end{array}
$$

Any rational number $\frac{n}{m}$ can be found in the row for n and the column for m. For example, $\frac{2}{3}$ (circled above) is in the row for 2 (the third row, since the first row is for 0) and the column for 3 (the third column). In fact, every rational number appears multiple times in the grid (infinitely many times, in fact). For example, the rational number $\frac{1}{2}$, which occurs in the second row, second column, is the same as the rational number $\frac{2}{4}$, which occurs in the third row, fourth column. (It’s also the same as $\frac{3}{6}, \frac{4}{8}, \frac{5}{10} \dots$.)

Our goal is to find a way to line up the naturals with the rationals—to find a one-to-one function, f, with domain $\mathbb{N}$ and range $\mathbb{Q}$. Since each rational number appears in the grid, all we need to do is go through all of the (infinitely many!) points on the grid, one by one, and count off a corresponding natural number for each; we’ll then let our function f map the natural numbers we count off to the rational numbers that appear at the corresponding points on the grid. Let’s start at the top left of the grid, and count off the first natural number, 0. So we’ll have f map 0 to the rational number at the top left of the grid, namely, $\frac{0}{1}$. That is, $f(0) = \frac{0}{1}$. We can depict this by labeling $\frac{0}{1}$ with the natural number we counted off, 0:

$$\begin{array}{cc|cccccc}
 & & \multicolumn{6}{c}{\text{denominators}} \\
 & & 1 & 2 & 3 & 4 & 5 & \dots \\
\cline{2-8}
 & 0 & \frac{0}{1}(0) & \frac{0}{2} & \frac{0}{3} & \frac{0}{4} & \frac{0}{5} & \dots \\
 & 1 & \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \dots \\
\text{numerators} & 2 & \frac{2}{1} & \frac{2}{2} & \frac{2}{3} & \frac{2}{4} & \frac{2}{5} & \dots \\
 & 3 & \frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \dots \\
 & 4 & \frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \dots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Next, ignoring a certain wrinkle which I’ll get to in a moment, let’s count off natural numbers for the rationals in the uppermost “ring” around the top left of the grid, in counterclockwise order, beginning at the left:

$$\begin{array}{cc|cccccc}
 & & \multicolumn{6}{c}{\text{denominators}} \\
 & & 1 & 2 & 3 & 4 & 5 & \cdots \\
\hline
 & 0 & \frac{0}{1}(0) & \frac{0}{2}(3) & \frac{0}{3} & \frac{0}{4} & \frac{0}{5} & \cdots \\
\text{numerators} & 1 & \frac{1}{1}(1) & \frac{1}{2}(2) & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\
 & 2 & \frac{2}{1} & \frac{2}{2} & \frac{2}{3} & \frac{2}{4} & \frac{2}{5} & \cdots \\
 & 3 & \frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \cdots \\
 & 4 & \frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Then (continuing to ignore the wrinkle) let's count off the next ring of numbers, again in counterclockwise order beginning at the left:

$$\begin{array}{cc|cccccc}
 & & \multicolumn{6}{c}{\text{denominators}} \\
 & & 1 & 2 & 3 & 4 & 5 & \cdots \\
\hline
 & 0 & \frac{0}{1}(0) & \frac{0}{2}(3) & \frac{0}{3}(8) & \frac{0}{4} & \frac{0}{5} & \cdots \\
\text{numerators} & 1 & \frac{1}{1}(1) & \frac{1}{2}(2) & \frac{1}{3}(7) & \frac{1}{4} & \frac{1}{5} & \cdots \\
 & 2 & \frac{2}{1}(4) & \frac{2}{2}(5) & \frac{2}{3}(6) & \frac{2}{4} & \frac{2}{5} & \cdots \\
 & 3 & \frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \cdots \\
 & 4 & \frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \cdots \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

And so on infinitely. For each new ring, we begin at the left, and move through the ring counterclockwise, continuing to count off natural numbers.

Every point on the grid will eventually be reached by one of these increasingly large (but always finite) rings. Since every rational number appears on the grid, every rational number eventually gets labeled with a natural number. So the range of our function f is the entirety of $\mathbb{Q}$! There are two tricks that make this work. First, even though the rational numbers are dense, they can be laid out in a discrete grid. Second, even though the grid is two dimensional and the natural numbers are only one-dimensional, there is a way to cover the whole grid with naturals since there is a "one-dimensional" path that covers the entire grid: the path along the expanding rings.

The wrinkle is that this procedure, as we've laid it out so far, doesn't deliver a one-to-one function, because rational numbers appear multiple times in the grid. For example, given our definition, f maps 0 to $\frac{0}{1}$ and 3 to $\frac{0}{2}$. But $\frac{0}{2}$ is the same rational number as $\frac{0}{1}$—namely, 0—so f isn't one-to-one. (f also maps 8 to 0; and it maps both 1 and 5 to 1, etc.) But it's easy to modify the procedure to fix this problem. In our trek through the rings, whenever we hit a rational number that we've already encountered, let's now simply skip it, and go on to the next rational number on the trek. Thus, the new diagram looks as follows (the skipped rational numbers are struck out):

denominators

		1	2	3	4	5	…
	0	$\frac{0}{1}$(0)	$\cancel{\frac{0}{2}}$	$\cancel{\frac{0}{3}}$	$\cancel{\frac{0}{4}}$	$\cancel{\frac{0}{5}}$	…
numerators	1	$\frac{1}{1}$(1)	$\frac{1}{2}$(2)	$\frac{1}{3}$(5)	$\frac{1}{4}$(9)	$\frac{1}{5}$(15)	…
	2	$\frac{2}{1}$(3)	$\cancel{\frac{2}{2}}$	$\frac{2}{3}$(4)	$\cancel{\frac{2}{4}}$	$\frac{2}{5}$(14)	…
	3	$\frac{3}{1}$(6)	$\frac{3}{2}$(7)	$\cancel{\frac{3}{3}}$	$\frac{3}{4}$(8)	$\frac{3}{5}$(13)	…
	4	$\frac{4}{1}$(10)	$\cancel{\frac{4}{2}}$	$\frac{4}{3}$(11)	$\cancel{\frac{4}{4}}$	$\frac{4}{5}$(12)	…
	⋮	⋮	⋮	⋮	⋮	⋮	⋱

We've now got our desired function f: it is the function that maps each natural number to the rational number in the grid labelled by that natural number. (Notice, incidentally, that f could be displayed in this way instead:

$$\begin{array}{lccccccccccccccccc} n: & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & \ldots \\ f(n): & 0 & 1 & \frac{1}{2} & 2 & \frac{2}{3} & \frac{1}{3} & 3 & \frac{3}{2} & \frac{3}{4} & \frac{1}{4} & 4 & \frac{4}{3} & \frac{4}{5} & \frac{3}{5} & \frac{2}{5} & \frac{1}{5} & \ldots \end{array}$$

This is just a different picture of the same function.) Since each rational number is labeled by some natural number, f's range is $\mathbb{Q}$. f's domain is clearly $\mathbb{N}$. And f is clearly one-to-one (since our procedure skips previously encountered rational numbers). So f is our desired function; $\mathbb{N}$ and $\mathbb{Q}$ are the same size.

If even a dense set like $\mathbb{Q}$ is no bigger than $\mathbb{N}$, are *all* infinite sets the same size? The answer is in fact no. Some infinite sets are bigger than $\mathbb{N}$; there are different sizes of infinity.

One such set is the set of real numbers. Real numbers are numbers that can be represented by decimals. All rational numbers are real numbers; and their decimal representations either terminate or eventually repeat in some infinitely recurring pattern. (For example, $\frac{1}{3}$ has the repeating decimal representation $0.3333\ldots$; $\frac{7}{4}$ has the terminating decimal representation 1.75.) But some real numbers are not rational numbers. These are the real numbers with decimal representations that never repeat. One example is the real number π, whose decimal representation begins: $3.14159\ldots$.

We'll prove that there are more real than natural numbers by proving that there are more real numbers between 0 and 1 than there are natural numbers. Let R be the set of real numbers in this interval. Now, consider the function f which maps each natural number n to $\frac{1}{n+2}$. This is a one-to-one function whose domain is $\mathbb{N}$ and whose range is $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$. But this latter set is a subset of R. So R is *at least* as big as $\mathbb{N}$. So all we need to do is show that R is not the same size as $\mathbb{N}$. And we can do this by showing that the assumption that $\mathbb{N}$ and R *are* the same size would lead to a contradiction.

So, suppose that $\mathbb{N}$ and R are equinumerous. Given the definition of equinumerosity, there must exist some one-to-one function, f, whose domain is $\mathbb{N}$ and whose range is R. We can represent f on a grid as follows:

$$\begin{array}{rcccccc}
f(0) = & 0 & . & a_{0,0} & a_{0,1} & a_{0,2} & \cdots \\
f(1) = & 0 & . & a_{1,0} & a_{1,1} & a_{1,2} & \cdots \\
f(2) = & 0 & . & a_{2,0} & a_{2,1} & a_{2,2} & \cdots \\
\vdots & \vdots & & \vdots & \vdots & \vdots & \ddots
\end{array}$$

The grid represents the real numbers in the range of f by their decimal representations.[12] The a's are the digits in these decimal representations. For any natural number i, $f(i)$ is represented as the decimal $0.a_{i,0}a_{i,1}a_{i,2}\ldots$. Thus $a_{i,j}$ is the $(j+1)^{\text{st}}$ digit in the decimal representation of $f(i)$. Consider $f(2)$, for example. If $f(2)$ happens to be the real number $0.2562894\ldots$, then $a_{2,0} = 2$, $a_{2,1} = 5$, $a_{2,2} = 6$, $a_{2,3} = 2$, and so on.

[12] If a decimal representation terminates, we can think of it as nevertheless being infinite: there are infinitely many zeros after the termination point.

The right hand part of the grid (everything except the column beginning with "$f(0) =$") is a list of real numbers. The first real number on this list is $0.a_{0,0}a_{1,1}a_{0,2}\ldots$, the second is $0.a_{1,0}a_{1,1}a_{1,2}\ldots$, the third is $0.a_{2,0}a_{2,1}a_{2,2}\ldots$, and so on. The real numbers in this list, in fact, comprise the range of f. But we have supposed, remember, that the range of f is the entirety of R. Thus, we have an important consequence of our supposition: this list is a *complete* list of R. That is, every member of R occurs somewhere on the list, as the decimal $0.a_{i,0}a_{i,1}a_{i,2}\ldots$, for some natural number i.

But in fact, we can show that this *can't* be a complete list of R, by showing that there is at least one real number between 0 and 1 that does not appear on the list. We're going to do this in a crafty way: we'll look at the grid above, and construct our real number as a function of the grid in such a way that it's guaranteed not to be anywhere on the list.

I'll call the real number I'm after "d"; to specify d, I'm going to specify its decimal representation $0.d_0d_1d_2\ldots$. Here is my definition of the j^{th} digit in this decimal representation:

$$d_j = \begin{cases} 6 & \text{if } a_{j,j} = 5 \\ 5 & \text{otherwise} \end{cases}$$

The "$a_{j,j}$"s refer to the grid depicting f above; thus, what real number d we have defined depends on the nature of the grid, and thus on the nature of the function f.

To get a handle on what's going on here, think about it geometrically. Consider the digits on the following *diagonal line* in the grid:

$$\begin{array}{cccccccc} f(0)= & 0 & . & \boxed{a_{0,0}} & a_{0,1} & a_{0,2} & \cdots \\ f(1)= & 0 & . & a_{1,0} & \boxed{a_{1,1}} & a_{1,2} & \cdots \\ f(2)= & 0 & . & a_{2,0} & a_{2,1} & \boxed{a_{2,2}} & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots \end{array}$$

To these diagonal digits, there corresponds a real number: $0.a_{0,0}a_{1,1}a_{2,2}\ldots$. Call this real number a. What we did to arrive at our number d (so-called because we are giving a "diagonal argument") was to begin with a's decimal representation and change each of its digits. We changed each of its digits to 5, except when the digit was already 5, in which case we changed it to 6.

We now approach the punch line. d's definition insures that it cannot be anywhere on the list. Let $f(i)$ be *any* member of the list. We can prove that d and $f(i)$ are not the same number. If they were, then their decimal representations $0.d_0d_1d_2\ldots$ and $0.a_{i,0}a_{i,1}a_{i,2}\ldots$ would also be the same. So each digit d_j in d's decimal representation would equal its corresponding digit $a_{i,j}$ in $f(i)$'s decimal representation. But this can't be. There is one place in particular where the digits must differ: the i^{th} place. d_i is defined to be 6 if $a_{i,i}$ is 5, and defined to be 5 if $a_{i,i}$ is not 5. Thus, d_i is not the same digit as $a_{i,i}$. So d's decimal representation differs in at least one place from $f(i)$'s decimal representation; so d is different from $f(i)$. But $f(i)$ was an arbitrarily chosen member of the list. Thus we have our conclusion: d isn't anywhere on the list. But d is a real number between 0 and 1. So if our initial assumption that the range of f is all of R were correct, d would have to be on the list. So that initial assumption was false, and we've completed our argument: it's impossible for there to be a one-to-one function whose domain is $\mathbb{N}$ and whose range is all of R. Even though $\mathbb{N}$ and R are both infinite sets, R is a bigger infinite set.

To grasp the argument's final phase, think again in geometric terms. If d were on the list, its decimal representation would intersect the diagonal. Suppose, for instance, that d were $f(3)$:

$$\begin{array}{rlcccccccc}
 & f(0)= & 0 & . & \boxed{a_{0,0}} & a_{0,1} & a_{0,2} & a_{0,3} & a_{0,4} & \cdots \\
 & f(1)= & 0 & . & a_{1,0} & \boxed{a_{1,1}} & a_{1,2} & a_{1,3} & a_{1,4} & \cdots \\
 & f(2)= & 0 & . & a_{2,0} & a_{2,1} & \boxed{a_{2,2}} & a_{2,3} & a_{2,4} & \cdots \\
d= & f(3)= & 0 & . & \boxed{a_{3,0}} & \boxed{a_{3,1}} & \boxed{a_{3,2}} & \boxed{\boxed{a_{3,3}}} & \boxed{a_{3,4}} & \cdots \\
 & f(4)= & 0 & . & a_{4,0} & a_{4,1} & a_{4,2} & a_{4,3} & \boxed{a_{4,4}} & \cdots \\
 & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Then, given d's definition, its decimal representation would be guaranteed to differ from the diagonal series in its fourth digit, the point of intersection.

It's natural to voice the following misgiving about the argument: "if d was left off the list, then why can't you just add it in? You could add it in at the beginning, bumping all the remaining members of the list down one slot to make room for it":

$$\begin{tikzcd}
\text{initial list} & \text{make room for } d & \text{new list} \\
f(0) & \arrow[d] & d \\
f(1) & f(0) \arrow[d] & f(0) \\
f(2) & f(1) \arrow[d] & f(1) \\
\vdots & f(2) \arrow[d] & f(2) \\
\vdots & \vdots & \vdots
\end{tikzcd}$$

Natural as it is, the misgiving is misguided. It's true that, given any list, one could add d to that list using the method described. But this fact is irrelevant to the argument. The argument wasn't that there is some *unlistable real number*, d—some real number d that is somehow prevented from occurring in the range of any one-to-one function whose domain is $\mathbb{N}$. That would be absurd. The argument was rather that no one list can be complete; any list (i.e., any one-to-one function whose domain is $\mathbb{N}$) will leave out some real numbers. The left-out real numbers can appear on *other* lists, but that's beside the point. Compare: if a thousand people show up to eat at a small restaurant, many people will be left out. That's not to say that any individual person is incapable of entering; it's just to say that not everyone can enter at once. No matter who enters, others will be left out in the cold.

Exercise 1.4* For any set, A, the powerset of A is defined as the set of all A's subsets. Write out the definition of the powerset of A in the "$\{u : \dots\}$" notation. Write out the powerset of $\{2,4,6\}$ in the braces notation (the one where you list each member of the set).

Exercise 1.5* Is $\mathbb{N}$ equinumerous with the set $\mathbb{Z}$ of all integers, negative, positive, and zero: $\{\cdots -3,-2,-1,0,1,2,3,\dots\}$?

Chapter 2

Propositional Logic

We begin with the simplest logic commonly studied: propositional logic. Despite its simplicity, it has great power and beauty.

2.1 Grammar of PL

We're going to approach propositional logic by studying a formal language. And the first step in the study of a formal language is always to rigorously define the language's grammar.

If all you want to do is to use and understand the language of logic, you needn't be so careful about grammar. For even without a precisely formulated grammar, you can intuitively recognize that things like this make sense:

$$P \rightarrow Q$$
$$R \wedge (\sim S \leftrightarrow P)$$

whereas things like this do not:

$$\rightarrow PQR\sim$$
$$(P\sim Q\sim(\vee$$
$$P \oplus Q$$

But to make any headway in metalogic, we will need more than an intuitive understanding of what makes sense and what does not. We will need a precise definition that has the consequence that only the strings of symbols in the first group "make sense".

Grammatical strings of symbols (i.e., ones that "make sense") are called *well-formed formulas*, or "formulas" or "wffs" for short. We define these by first carefully defining exactly which symbols are allowed to occur in wffs (the "primitive vocabulary"), and second, carefully defining exactly which strings of these symbols count as wffs. Here is the official definition; I'll explain what it means in a moment:

Primitive vocabulary:

- Connectives:[1] $\rightarrow, \sim$
- Sentence letters: $P, Q, R\dots$, with or without numerical subscripts
- Parentheses: (,)

Definition of wff:

i) Every sentence letter is a PL-wff

ii) If ϕ and ψ are PL-wffs then $(\phi\rightarrow\psi)$ and $\sim\phi$ are also PL-wffs

iii) Only strings that can be shown to be PL-wffs using i) and ii) are PL-wffs

(We allow numerical subscripts on sentence letters so that we don't run out when constructing increasingly complex formulas. Since $P_1, P_2, P_3\dots$ are all sentence letters, we have infinitely many to choose from.)

We will be discussing a number of different logical systems throughout this book, with differing grammars. What we have defined here is the notion of a wff for one particular language, the language of PL. So strictly, we should speak of *PL-wffs*, as the official definition does. But usually I'll just say "wff" if there is no danger of ambiguity.

Here is how the definition works. Its core is clauses i) and ii) (they're sometimes called the formation rules). Clause i) says that if you write down a sentence letter on its own, that counts as a wff. So, for example, the sentence letter P, all by itself, is a wff. (So is Q, so is P_{147}, and so on. Sentence letters are often called "atomic" wffs, because they're not made up of smaller wffs.) Next, clause ii) tells us how to build complex wffs from smaller wffs. It tells us that we can do this in two ways. First, it says that if we already have a wff, then we can put a $\sim$ in front of it to get another wff. (The resulting wff is often called a

[1] Some books use $\supset$ instead of $\rightarrow$, or $\neg$ instead of $\sim$. Other common symbols include & or $\cdot$ for conjunction, | for disjunction, and $\equiv$ for the biconditional.

"negation".) For example, since P is a wff (we just used clause i) to establish this), then $\sim P$ is also a wff. Second, clause ii) says that if we already have two wffs, then we can put an → between them, enclose the whole thing in parentheses, and we get another wff. (The resulting wff is often called a "conditional", whose "antecedent" is the wff before the → and whose "consequent" is the wff after the →.) For example, since we know that Q is a wff (clause i)), and that $\sim P$ is a wff (we just showed this a moment ago), we know that $(Q \rightarrow \sim P)$ is also a wff. This process can continue. For example, we could put an → between the wff we just constructed and R (which we know to be a wff from clause i)) to construct another wff: $((Q \rightarrow \sim P) \rightarrow R)$. By iterating this procedure, we can demonstrate the wffhood of arbitrarily complex strings.

Why the greek letters in clause ii)? Well, it wouldn't be right to phrase it, for example, in the following way: "if P and Q are wffs, then $\sim P$ and $(P \rightarrow Q)$ are also wffs". That would be too narrow, for it would apply only in the case of the sentence letters P and Q. It wouldn't apply to any other sentence letters (it wouldn't tell us that $\sim R$ is a wff, for example), nor would it allow us to construct negations and conditionals from complex wffs (it wouldn't tell us that $(P \rightarrow \sim Q)$ is a wff). We want to say that for *any* wff (not just P), if you put a ∼ in front of it you get another wff; and for *any* two wffs (not just P and Q), if you put an → between them (and enclose the result in parentheses) you get another wff. That's why we use the *metalinguistic variables* "ϕ" and "ψ".[2] The practice of using variables to express generality is familiar; we can say, for example, "for any integer n, if n is even, then $n + 2$ is even as well". Just as "n" here is a variable for numbers, metalinguistic variables are variables for linguistic items. (We call them *meta*linguistic because they are variables we use in our metalanguage, in order to talk generally about the object language, which is in this case the formal language of propositional logic.)

What's the point of clause iii)? Clauses i) and ii) provide only sufficient conditions for being a wff, and therefore do not on their own exclude nonsense combinations of primitive vocabulary like $P \sim Q \sim R$, or even strings like $P \oplus Q$ that include disallowed symbols. Clause iii) rules these strings out, since there is no way to build up either of these strings from clauses i) and ii), in the way that we built up the wff $(\sim P \rightarrow (P \rightarrow Q))$.

Notice an interesting feature of this definition: the very expression we are trying to define, 'wff', appears on the right hand side of clause ii) of the definition. In a sense, we are using the expression 'wff' in its own definition. But

[2] Strictly speaking clause iii) ought to be phrased using corner quotes; see exercise 1.2b.

this "circularity" is benign, because the definition is *recursive*. A recursive (or "inductive") definition of a concept F contains a circular-seeming clause, often called the "inductive" clause, which specifies that *if* such-and-such objects are F, then so-and-so objects are also F. But a recursive definition also contains a "base clause", which specifies noncircularly that certain objects are F. Even though the inductive clause rests the status of certain objects as being Fs on whether certain other objects are Fs (whose status as Fs might in turn depend on the status of still other objects…), this eventually traces back to the base clause, which secures F-hood all on its own. Thus, recursive definitions are anchored by their base clauses; that's what distinguishes them from viciously circular definitions. In the definition of wffs, clause i) is the base, and clause ii) is the inductive clause. The wffhood of the string of symbols $((P\rightarrow Q)\rightarrow\sim R)$, for example, rests on the wffhood of $(P\rightarrow Q)$ and of $\sim R$ by clause ii); and the wffhood of these, in turn, rests on the wffhood of P, Q and R, again by clause ii). But the wffhood of P, Q, and R doesn't rest on the wffhood of anything else; clause i) specifies directly that all sentence letters are wffs.

What happened to $\wedge$, $\vee$, and $\leftrightarrow$? The only connectives in our primitive vocabulary are $\rightarrow$ and $\sim$; expressions like $P\wedge Q$, $P\vee Q$, and $P\leftrightarrow Q$ therefore do not officially count as wffs. But we can still use $\wedge$, $\vee$, and $\leftrightarrow$ unofficially, since we can define those connectives in terms of $\sim$ and $\rightarrow$:

Definitions of $\wedge$, $\vee$, and $\leftrightarrow$:

- "$\phi\wedge\psi$" is short for "$\sim(\phi\rightarrow\sim\psi)$"
- "$\phi\vee\psi$" is short for "$\sim\phi\rightarrow\psi$"
- "$\phi\leftrightarrow\psi$" is short for "$(\phi\rightarrow\psi)\wedge(\psi\rightarrow\phi)$" (which is in turn short for "$\sim((\phi\rightarrow\psi)\rightarrow\sim(\psi\rightarrow\phi))$")

So, whenever we subsequently write down an expression that includes one of the defined connectives, we can regard it as being short for an expression that includes only the official connectives, $\sim$ and $\rightarrow$. (Why did we choose these particular definitions? We'll show below that they generate the usual truth conditions for $\wedge$, $\vee$, and $\leftrightarrow$.)

Our choice to begin with $\rightarrow$ and $\sim$ as our official connectives was somewhat arbitrary. We could have started with $\sim$ and $\wedge$, and defined the others as follows:

- "$\phi\vee\psi$" is short for "$\sim(\sim\phi\wedge\sim\psi)$"
- "$\phi\rightarrow\psi$" is short for "$\sim(\phi\wedge\sim\psi)$"

- "$\phi \leftrightarrow \psi$" is short for "$(\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$"

And other alternate choices are possible. (Why did we choose only a small number of primitive connectives, rather than including all of the usual connectives? Because, as we will see, it makes metalogic easier.)

The definition of wff requires conditionals to have outer parentheses. $P \rightarrow Q$, for example, is officially not a wff; one must write $(P \rightarrow Q)$. But informally, I'll often omit those outer parentheses. And I'll sometimes write square brackets instead of the official round ones (for example, "$[(P \rightarrow Q) \rightarrow R] \rightarrow P$") to improve readability.

2.2 The semantic approach to logic

In the next section I will introduce a "semantics" for propositional logic, and formal representations of logical truth and logical consequence of the semantic (model-theoretic) variety (recall section 1.5).

On the semantic conception, logical consequence amounts to: truth-preservation in virtue of the meanings of the logical constants. This slogan isn't perfectly clear, but it does lead to a clearer thought: suppose we keep the meanings of an argument's logical constants fixed, but vary *everything else*. If the argument remains truth-preserving no matter how we vary everything else, then it would seem to preserve truth "in virtue of" the meanings of its logical constants. But what is to be included in "everything else"?

Here is an attractive picture of truth and meaning. The truth of a sentence is determined by two factors, *meaning* and *the world*. A sentence's meaning determines the conditions under which its true—the ways the world would have to be, in order for that sentence to be true. If the world *is* one of the ways picked out by the sentence's truth conditions, then the sentence is true; otherwise, not. Furthermore, a sentence's meaning is typically determined by the meanings of its parts—both its logical constants and its nonlogical expressions. So: three elements determine whether a sentence is true: the world, the meanings of its nonlogical expressions, and the meanings of its logical constants.[3]

Now we can say what "everything else" means. Since we're holding constant the third element (the meanings of logical constants), varying everything else means varying the first two elements. The clearer thought about logical consequence, then, is that if an argument remains truth-preserving no matter

[3] And also a fourth element: its syntax. We hold this constant as well.

how we vary i) the world, and ii) the meanings of nonlogical expressions, then its premises logically imply its conclusion.

To turn this clearer, but still not perfectly clear, thought into a formal approach, we need to do two things. First, we need mathematical representations—I'll call them *configurations*—of variations of types i) and ii). A configuration is a mathematical representation, both of the world and of the meanings of nonlogical expressions. Second, we need to define the conditions under which a sentence of the formal language in question is *true in* one of these configurations. When we've done both things, we'll have a semantics for our formal language.

One thing such a semantics is good for, is giving a formalization, of the semantic variety, of the notions of logical consequence and logical truth. This formalization represents one formula as being a logical consequence of others iff it is true in any configuration in which the latter formulas are true, and represents a formula as being a logical truth iff it is true in all configurations.

But a semantics for a formal language is good for something else as well. Defining configurations, and truth-in-a-configuration, can shed light on meaning in natural and other interpreted languages.

Philosophers disagree over how to understand the notion of meaning in general. But meaning surely has *something* to do with truth conditions, as in the attractive picture above. If so, a formal semantics can shed light on meaning, if the ways in which configurations render formal sentences true and false are parallel to the ways in which the real world plus the meanings of words render corresponding interpreted sentences true and false. Expressions in formal languages are typically intended to represent bits of interpreted languages. The PL logical constant $\sim$, for example, represents the English logical constant 'not'; the sentence letters represent English declarative sentences, and so on. Part of specifying a configuration will be specifying what the nonlogical expressions mean in that configuration. And the definition of truth-in-a-configuration will be constructed so that the contributions of the symbolic logical constants to truth-conditions will mirror the contributions to truth conditions of the logical constants that they represent.

2.3 Semantics of propositional logic

Our semantics for propositional logic is really just a more rigorous version of the method of truth tables from introductory logic books. What a truth table does is depict how the truth value of a given formula is determined by the truth values of its sentence letters, for *each* possible combination of truth values for its sentence letters. To do this nonpictorially, we need to define a notion corresponding to "a possible combination of truth values for sentence letters":

Definition of interpretation: A PL-interpretation is a function $\mathscr{I}$, that assigns to each sentence letter either 1 or 0

The numbers 1 and 0 are our truth values. (Sometimes the letters 'T' and 'F' are used instead.) So an interpretation assigns truth values to sentence letters. Instead of saying "let P be false, and Q be true", we can say: let $\mathscr{I}$ be an interpretation such that $\mathscr{I}(P)=0$ and $\mathscr{I}(Q)=1$. (As with the notion of a wff, we will have different definitions of interpretations for different logical systems, so strictly we must speak of *PL-interpretations*. But usually it will be fine to speak simply of interpretations when it's clear which system is at issue.)

An interpretation assigns a truth value to each of the infinitely many sentence letters. To picture one such interpretation we could begin as follows:

$$\mathscr{I}(P)=1$$
$$\mathscr{I}(Q)=1$$
$$\mathscr{I}(R)=0$$
$$\mathscr{I}(P_1)=0$$
$$\mathscr{I}(P_2)=1$$

but since there are infinitely many sentence letters, the picture could not be completed. And this is just one interpretation among infinitely many; any other combination of assigned 1s and 0s to the infinitely many sentence letters counts as a new interpretation.

Once we settle what truth values a given interpretation assigns to the sentence letters, the truth values of complex sentences containing those sentence letters are thereby fixed. The usual, informal, method for showing exactly how those truth values are fixed is by giving truth tables for each connective. The standard truth tables for the → and ~ are the following:[4]

$\rightarrow$	1	0
1	1	0
0	1	1

$\sim$	
1	0
0	1

What we will do, instead, is write out a formal definition of a function—the *valuation* function—that assigns truth values to complex sentences as a function of the truth values of their sentence letters—i.e., as a function of a given intepretation $\mathscr{I}$. But the idea is the same as the truth tables: truth tables are really just pictures of the definition of a valuation function.

Definition of valuation: For any PL-interpretation, $\mathscr{I}$, the PL-valuation for $\mathscr{I}$, $V_{\mathscr{I}}$, is defined as the function that assigns to each wff either 1 or 0, and which is such that, for any sentence letter α and any wffs ϕ and ψ:

$$\begin{aligned} V_{\mathscr{I}}(\alpha) &= \mathscr{I}(\alpha) \\ V_{\mathscr{I}}(\phi \rightarrow \psi) &= 1 \text{ iff either } V_{\mathscr{I}}(\phi) = 0 \text{ or } V_{\mathscr{I}}(\psi) = 1 \\ V_{\mathscr{I}}({\sim}\phi) &= 1 \text{ iff } V_{\mathscr{I}}(\phi) = 0 \end{aligned}$$

Intuitively: we begin by choosing an interpretation function, which fixes the truth values for sentence letters. Then the valuation function assigns corresponding truth values to complex sentences depending on what connectives they're built up from: a negation is true iff the negated formula is false, and a conditional is true when its antecedent is false or its consequent is true.

We have here another recursive definition: the valuation function's values for complex formulas are determined by its values for smaller formulas; and this procedure bottoms out in the values for sentence letters, which are determined directly by the interpretation function $\mathscr{I}$.

Notice how the definition of the valuation function contains the English logical connectives 'either…or', and 'iff'. I used these English connectives rather than the logical connectives ∨ and ↔, because at that point I was *not*

[4] The → table, for example, shows what truth value $\phi \rightarrow \psi$ takes on depending on the truth values of its parts. Rows correspond to truth values for ϕ, columns to truth values for ψ. Thus, to ascertain the truth value of $\phi \rightarrow \psi$ when ϕ is 1 and ψ is 0, we look in the 1 row and the 0 column. The listed value there is 0—the conditional is false in this case. The ~ table has only one "input-column" and one "result-column" because ~ is a one-place connective.

writing down wffs of the language of study (in this case, the language of propositional logic). I was rather using sentences of English—our metalanguage, the informal language we're using to discuss the formal language of propositional logic—to construct my definition of the valuation function. My definition needed to employ the logical notions of disjunction and biconditionalization, the English words for which are 'either…or' and 'iff'.

One might again worry that something circular is going on. We defined the symbols for disjunction and biconditionalization, $\vee$ and $\leftrightarrow$, in terms of $\sim$ and $\rightarrow$ in section 2.1, and now we've defined the valuation function in terms of disjunction and biconditionalization. So haven't we given a circular definition of disjunction and biconditionalization? No. When we define the valuation function, we're not trying to *define* logical concepts such as negation, conjunction, disjunction, conditionalization, and biconditionalization, and so on, at all. Reductive definition of these very basic concepts is probably impossible (though one can define some of them in terms of the others). What we are doing is starting with the assumption that we *already* understand the logical concepts, and then using those concepts to provide a semantics for a formal language. This can be put in terms of object- and meta-language: we use metalanguage connectives, such as 'iff' and 'or', which we simply take ourselves to understand, to provide a semantics for the object language connectives $\sim$ and $\rightarrow$.

An elementary fact will be important in what follows: for every wff ϕ and every PL-interpretation $\mathscr{I}$, $V_{\mathscr{I}}(\phi)$ is either 0 or 1, but not both.[5] Equivalently: a formula has one of the truth values iff it lacks the other. That this is a fact is built into the definition of the valuation function for PL. First of all, $V_{\mathscr{I}}$ is defined as a *function*, and so it can't assign *both* the number 0 and the number 1 to a wff. And second, $V_{\mathscr{I}}$ is defined as a function that *assigns either* 1 *or* 0 *to each wff* (thus, in the case of the second and third clauses, if a complex wff fails the condition for getting assigned 1, it automatically gets assigned 0.)

Back to the definition of the valuation function. The definition applies only to official wffs, which can contain only the primitive connectives $\rightarrow$ and $\sim$. But sentences containing $\wedge$, $\vee$, and $\leftrightarrow$ are abbreviations for official wffs, and are therefore indirectly governed by the definition. In fact, given the abbreviations defined in section 2.1, we can show that the definition assigns the intuitively

[5] This fact won't hold for all the valuation functions we'll consider in this book; in chapter 3 we will consider "trivalent" semantic systems in which some formulas are assigned neither 1 nor 0.

correct truth values to sentences containing ∧, ∨, and ↔. In particular, we can show that for any PL-interpretation $\mathscr{I}$, and any wffs ψ and χ,

$$\begin{aligned} \mathrm{V}_{\mathscr{I}}(\psi\wedge\chi)=1 &\text{ iff } \mathrm{V}_{\mathscr{I}}(\psi)=1 \text{ and } \mathrm{V}_{\mathscr{I}}(\chi)=1\\ \mathrm{V}_{\mathscr{I}}(\psi\vee\chi)=1 &\text{ iff either } \mathrm{V}_{\mathscr{I}}(\psi)=1 \text{ or } \mathrm{V}_{\mathscr{I}}(\chi)=1\\ \mathrm{V}_{\mathscr{I}}(\psi\leftrightarrow\chi)=1 &\text{ iff } \mathrm{V}_{\mathscr{I}}(\psi)=\mathrm{V}_{\mathscr{I}}(\chi) \end{aligned}$$

I'll show that the first statement is true here; the others are exercises for the reader. I'll write out this proof in excessive detail, to make it clear exactly how the reasoning works.

Example 2.1: Proof that ∧ gets the right truth condition. We are to show that for every wffs ψ and χ, and any PL-interpretation $\mathscr{I}$, $\mathrm{V}_{\mathscr{I}}(\psi\wedge\chi)=1$ iff $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$. So, let ψ and χ be any wffs, and let $\mathscr{I}$ be any PL-interpretation; we must show that: $\mathrm{V}_{\mathscr{I}}(\psi\wedge\chi)=1$ iff $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$. The expression $\psi\wedge\chi$ is an abbreviation for the expression $\sim(\psi\rightarrow\sim\chi)$. So what we must show is this: $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$ iff $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$.

Now, in order to show that a statement *A* holds iff a statement *B* holds, we must first show that if *A* holds, then *B* holds; then we must show that if *B* holds then *A* holds. So, first we must establish that if $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$, then $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$. So, we begin by *assuming* that $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$, and we then attempt to show that $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$. Well, since $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$, by definition of the valuation function, clause for ∼, we know that $\mathrm{V}_{\mathscr{I}}(\psi\rightarrow\sim\chi)=0$. Now, we earlier noted the principle that a wff has one of the two truth values iff it lacks the other; thus, $\mathrm{V}_{\mathscr{I}}(\psi\rightarrow\sim\chi)$ is *not* 1. (Henceforth I won't mention it when I make use of this principle.) But then, by the clause in the definition of $\mathrm{V}_{\mathscr{I}}$ for the →, we know that it's not the case that: either $\mathrm{V}_{\mathscr{I}}(\psi)=0$ or $\mathrm{V}_{\mathscr{I}}(\sim\chi)=1$. So, $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\sim\chi)=0$. From the latter, by the clause for ∼, we know that $\mathrm{V}_{\mathscr{I}}(\chi)=1$. So now we have what we wanted: $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$.

Next we must show that if $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$, then $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$. This is sort of like undoing the previous half. Suppose that $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\chi)=1$. Since $\mathrm{V}_{\mathscr{I}}(\chi)=1$, by the clause for ∼, $\mathrm{V}_{\mathscr{I}}(\sim\chi)=0$; but now since $\mathrm{V}_{\mathscr{I}}(\psi)=1$ and $\mathrm{V}_{\mathscr{I}}(\sim\chi)=0$, by the clause for → we know that $\mathrm{V}_{\mathscr{I}}(\psi\rightarrow\sim\chi)=0$; then by the clause for ∼, we know that $\mathrm{V}_{\mathscr{I}}(\sim(\psi\rightarrow\sim\chi))=1$, which is what we were trying to show. ■

Example 2.1 is the first of many *metalogic proofs* we will be constructing in this book. (The symbol ■ marks the end of such a proof.) It is an informal argument, phrased in the metalanguage, which establishes a fact about a formal language. As noted in section 1.3, metalogic proofs must be distinguished from proofs in formal systems—from the derivations and truth trees of introductory logic, and from the axiomatic and sequent proofs we will introduce below. Although there are no explicit guidelines for how to present metalogic proofs, they are generally given in a style that is common within mathematics. Constructing such proofs can at first be difficult. I offer the following pointers. First, keep in mind exactly what you are trying to prove. (In your first few proofs, it might be a good idea to begin by writing down: "what I am trying to prove is…".) Second, keep in mind the definitions of all the relevant technical terms (the definition of $\psi\wedge\chi$, for instance.) Third, keep in mind exactly what you are *given*. (In the preceding, for example, the important bit of information you are given is the definition of the valuation function; that definition tells you the conditions under which valuation functions assign 1s and 0s to negations and conditionals.) Fourth, keep in mind the canonical methods for establishing claims of various forms. (For example, if you want to show that a certain claim holds for *every* two wffs, begin with "let ψ and χ be any wffs"; show that the claim holds for ψ and χ; and conclude that the claim holds for all pairs of wffs. If you want to establish something of the form "if A, then B", begin by saying "suppose A", go on to reason your way to "B", and conclude: "and so, if A then B." Often it can be helpful to reason by *reductio ad absurdum*: assume the opposite of the assertion you are trying to prove, reason your way to a contradiction, and conclude that the assertion is true since its opposite leads to contradiction.) Fifth: practice, practice, practice. As we progress, I'll gradually speed up the presentation of such proofs, omitting more and more details when they seem obvious. You should feel free to do the same; but it may be best to begin by constructing proofs very deliberately, so that later on you know exactly what details you are omitting.

Let's reflect on what we've done so far. We have defined the notion of a PL-interpretation, which assigns 1s and 0s to sentence letters of the formal language of propositional logic. And we have also defined, for any PL-interpretation, a corresponding PL-valuation function, which extends the interpretation's assignment of 1s and 0s to complex wffs of PL. Note that we have been informally speaking of these assignments as assignments of *truth values*. That's because the assignment of 1s and 0s to complex wffs mirrors the way complex natural language sentences get their truth values, as a function of the truth values of their parts. For example, the $\sim$ of propositional logic is supposed to represent the English phrase 'it is not the case that'. Accordingly, just as an English sentence "It is not the case that ϕ" is true iff ϕ is false, one of our valuation functions assigns 1 to $\sim\phi$ iff it assigns 0 to ϕ. But strictly, it's probably best not to think of wffs of our formal language as genuinely having truth values. They don't genuinely have meanings after all. Our assignments of 1 and 0 *represent* the having of truth values.

A semantics for a formal language, recall, defines two things: configurations and truth-in-a-configuration. In the propositional logic semantics we have laid out, the configurations are the interpretation functions. A configuration is supposed to represent a way for the world to be, plus the meanings of nonlogical expressions. The only nonlogical expressions in PL are the sentence letters; and, for the purposes of PL anyway, their meanings can be represented simply as truth-values. And once we've specified a truth-value for each sentence letter, we've already represented the world as much as we can in PL. Thus, PL-interpretations are appropriate configurations. As for truth-in-a-configuration, this is accomplished by the valuation functions. For any PL-interpretation, its corresponding valuation function specifies, for each complex wff, what truth value that wff has in that interpretation. Thus, for each wff (ϕ) and each configuration ($\mathscr{I}$), we have specified the truth value of that wff in that configuration ($V_{\mathscr{I}}(\phi)$).

Onward. We are now in a position to define the semantic versions of the notions of logical truth and logical consequence for propositional logic. The semantic notion of a logical truth is that of a *valid formula*:

Definition of validity: A wff ϕ is PL-valid iff for every PL-interpretation, $\mathscr{I}$, $V_{\mathscr{I}}(\phi) = 1$

We write "$\vDash_{PL} \phi$" for "ϕ is PL-valid". (When it's obvious which system we're talking about, we'll omit the subscript on $\vDash$.) The valid formulas of propositional logic are also called *tautologies*.

As for logical consequence, the semantic version of this notion is that of a single formula's being a *semantic consequence* of a set of formulas:

Definition of semantic consequence: A wff ϕ is a PL-semantic consequence of a set of wffs Γ iff for every PL-interpretation, $\mathscr{I}$, if $V_{\mathscr{I}}(\gamma) = 1$ for each γ such that $\gamma \in \Gamma$, then $V_{\mathscr{I}}(\phi) = 1$

That is, ϕ is a PL-semantic consequence of Γ iff ϕ is true whenever each member of Γ is true. We write "$\Gamma \vDash_{PL} \phi$" for "ϕ is a PL-semantic consequence of Γ". (As usual we'll often omit the "PL" subscript; and further, let's improve readability by writing "$\phi_1, \dots, \phi_n \vDash \psi$" instead of "$\{\phi_1, \dots, \phi_n\} \vDash \psi$". That is, let's drop the set braces when it's convenient to do so.)

A related concept is that of *semantic equivalence*. Formulas ϕ and ψ are said to be (PL-) semantically equivalent iff each (PL-) semantically implies the other. For example, $\phi\rightarrow\psi$ and $\sim\psi\rightarrow\sim\phi$ are semantically equivalent. Notice that we could just as well have worded the definition thus: semantically equivalent formulas are those that have exactly the same truth value in every interpretation. Thus, there is a sense in which semantically equivalent formulas "say the same thing": they have the same truth-conditional content.

Just as it's probably best not to think of sentences of our formal language as genuinely having truth values, it's probably best not to think of them as genuinely being logically true or genuinely standing in the relation of logical consequence. The notions we have just defined, of PL-validity and PL-semantic-consequence, are just formal representations of logical truth and logical consequence (semantically conceived). Indeed, the definitions we have given are best thought of as representing, rather than really being, a semantics. Further, when we get to formal provability, the definitions we will give are probably best thought of as representing facts about provability, rather than themselves defining a kind of provability. But forgive me if I sometimes speak loosely as if formal sentences really do have these features, rather than just representing them.

By the way, we can now appreciate why it was important to set up our grammar so carefully. The valuation function assigns truth values to complex formulas based on their form. One clause in its definition kicks in for atomic wffs, another clause kicks in for wffs of the form $\sim\phi$, and a third kicks in for wffs of the form $\phi\rightarrow\psi$. This works only if each wff has exactly one of these three forms; only a precise definition of wff guarantees this.

Exercise 2.1 Given the definitions of the defined symbols $\vee$ and $\leftrightarrow$, show that for any PL-interpretation, $\mathscr{I}$, and any wffs ψ and χ,

$$\begin{aligned} &V_{\mathscr{I}}(\psi\vee\chi) = 1 \text{ iff either } V_{\mathscr{I}}(\psi) = 1 \text{ or } V_{\mathscr{I}}(\chi) = 1 \\ &V_{\mathscr{I}}(\psi\leftrightarrow\chi) = 1 \text{ iff } V_{\mathscr{I}}(\psi) = V_{\mathscr{I}}(\chi) \end{aligned}$$

2.4 Establishing validity and invalidity in PL

Now that we have set up a semantics, we can establish semantic facts about particular wffs. For example:

Example 2.2: Proof that $\vDash_{\mathrm{PL}} (P\rightarrow Q)\rightarrow({\sim}Q\rightarrow{\sim}P)$. To show a wff to be PL-valid, we must show that it is true in every PL-interpretation. So, let $\mathscr{I}$ be any PL-interpretation, and suppose for reductio that $\mathrm{V}_{\mathscr{I}}((P\rightarrow Q)\rightarrow({\sim}Q\rightarrow{\sim}P)) = 0$. This assumption leads to a contradiction, as the following argument shows:

i) $\mathrm{V}_{\mathscr{I}}((P\rightarrow Q)\rightarrow({\sim}Q\rightarrow{\sim}P)) = 0$ (reductio assumption)

ii) So, by the definition of a valuation function, clause for the →, $\mathrm{V}_{\mathscr{I}}(P\rightarrow Q) = 1$ and…

iii) …$\mathrm{V}_{\mathscr{I}}({\sim}Q\rightarrow{\sim}P) = 0$

iv) Given iii), again by the clause for the →, $\mathrm{V}_{\mathscr{I}}({\sim}Q) = 1$ and …

v) …$\mathrm{V}_{\mathscr{I}}({\sim}P) = 0$

vi) Given iv), by the clause for the ~, $\mathrm{V}_{\mathscr{I}}(Q) = 0$.

vii) Similarly, v) tells us that $\mathrm{V}_{\mathscr{I}}(P) = 1$.

viii) From vii) and vi), by the clause for the → we know that $\mathrm{V}_{\mathscr{I}}(P\rightarrow Q) = 0$, which contradicts line ii).

Here again we have given a metalogic proof: an informal mathematical argument establishing a fact about one of our formal languages. (The conclusion of the argument was not sufficiently impressive to merit the ■ flourish at the end.) There is nothing special about the form that this argument took. One could just as well have established the fact that $\vDash_{\mathrm{PL}} (P\rightarrow Q)\rightarrow({\sim}Q\rightarrow{\sim}P)$ by constructing a truth table, as one does in introductory textbooks, for such a construction is in effect a pictorial metalogic proof that a certain formula is PL-valid.

Arguments establishing facts of semantic consequence are parallel (in this example we will proceed more briskly):

Example 2.3: Proof that $P\rightarrow(Q\rightarrow R) \vDash Q\rightarrow(P\rightarrow R)$. We must show that in any PL-interpretation in which $P\rightarrow(Q\rightarrow R)$ is true, $Q\rightarrow(P\rightarrow R)$ is true as well. Let $\mathscr{I}$ be any PL-interpretation; we then reason as follows:

i) Suppose for reductio that $V_{\mathscr{I}}(P\rightarrow(Q\rightarrow R)) = 1$ but…

ii) …$V_{\mathscr{I}}(Q\rightarrow(P\rightarrow R)) = 0$. (From now on we'll omit the subscripted $\mathscr{I}$.)

iii) line ii) tells us that $V(Q) = 1$ and $V(P\rightarrow R) = 0$, and hence that $V(R) = 0$. So $V(Q\rightarrow R) = 0$.

iv) Since $V(P\rightarrow R) = 0$ (line iii)), $V(P) = 1$. So then, by iii), $V(P\rightarrow(Q\rightarrow R)) = 0$. This contradicts i).

One can also establish facts of invalidity and failures of semantic consequence:

Example 2.4: Proof that $\nvDash ((P\wedge R)\rightarrow Q)\rightarrow(R\rightarrow Q)$. To be valid is to be true in all interpretations; so to be *invalid* (i.e., not valid) is to be false in at least one interpretation. So all we must do is find one interpretation in which this wff is false. Let $\mathscr{I}$ be an interpretation such that $\mathscr{I}(R) = 1$ and $\mathscr{I}(P) = \mathscr{I}(Q) = 0$. Then $V_{\mathscr{I}}(P\wedge R) = 0$ (example 2.1), so $V_{\mathscr{I}}((P\wedge R)\rightarrow Q) = 1$. But since $V_{\mathscr{I}}(R) = 1$ and $V_{\mathscr{I}}(Q) = 0$, $V_{\mathscr{I}}(R\rightarrow Q) = 0$. So $V_{\mathscr{I}}((P\wedge R)\rightarrow Q)\rightarrow(R\rightarrow Q)) = 0$

Example 2.5: Proof that $P\rightarrow R \nvDash (P\vee Q)\rightarrow R$. Consider a PL-interpretation in which P and R are false, and in which Q is true. $P\rightarrow R$ is then true (since its antecedent is false), but $P\vee Q$ is true (since Q is true—see exercise 2.1) while R is false, so $(P\vee Q)\rightarrow R$ is false.

I'll end this section by noting a certain fact about validity in propositional logic: it is mechanically "decidable". That is, a computer program could be written that is capable of telling, for any given formula, whether or not that formula is valid. The program would simply construct a complete truth table for the formula in question. To give a rigorous proof of this fact would take us too far afield, since we would need to give a rigorous definition of what counts as a computer program, but the point is intuitively clear.

Exercise 2.2 Establish each of the following facts:

a) $\vDash [P\wedge(Q\vee R)] \rightarrow [(P\wedge Q)\vee(P\wedge R)]$

b) $(P\leftrightarrow Q)\vee(R\leftrightarrow S) \nvDash P\vee R$

c) $\sim(P\wedge Q)$ and $\sim P\vee\sim Q$ are semantically equivalent.

2.7 Soundness of PL and proof by induction

Note: the next three sections are more difficult than the preceding sections, and may be skipped without much loss. If you decide to work through the more difficult sections dealing with metalogic later in the book (for example sections 6.5 and 6.6), you might first return to these sections.

In this chapter we have taken both a proof-theoretic and a semantic approach to propositional logic. In each case, we introduced formal notions of logical truth and logical consequence. For the semantic approach, these notions involved truth in PL-interpretations. For the proof-theoretic approach, we considered two formal definitions, one involving sequent proofs, the other involving axiomatic proofs.

An embarrassment of riches! We have multiple formal accounts of our logical notions. But in fact, it can be shown that *all three of our definitions yield* *exactly the same results*. Here I'll prove this just for the notion of a *theorem* (last line of an axiomatic proof) and the notion of a *valid* formula (true in all PL-interpretations). I'll do this by proving the following two statements:

Soundness of PL: Every PL-theorem is PL-valid

Completeness of PL: Every PL-valid wff is a PL-theorem

Soundness is pretty easy to prove; we'll do that in a moment. Completeness is harder; we'll prove that in section 2.9. Soundness and completeness together tell us that PL-validity and PL-theoremhood exactly coincide.

But first a short detour: we need to introduce a method of proof that is ubiquitous throughout metalogic (as well as mathematics generally), the method of induction. The basic idea, in its simplest form, is this. Suppose we have infinitely many objects lined up like this:

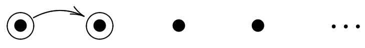
● ● ● ● ···

And suppose we want to show that each of these objects has a certain property. How to do it?

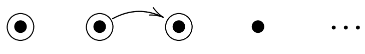
The method of induction directs us to proceed in two steps. First, show that the *first* object has the property:

⦿ ● ● ● ···

This is called the "base case" of the inductive proof. Next, show that quite generally, whenever one object in the line has the property, then the next must have the property as well. This is called the "inductive step" of the proof. The method of induction then says: if you've established those two things, you can go ahead and conclude that *all* the objects in the line have the property. Why is this conclusion justified? Well, since the first object has the property, the second object must have the property as well, given the inductive step:



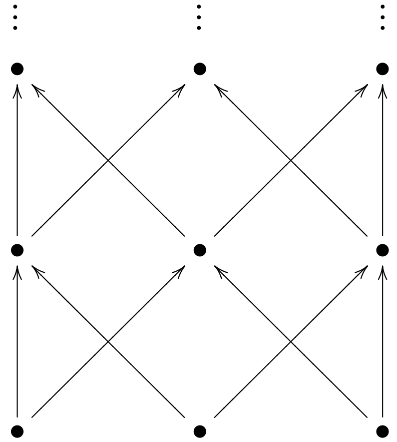
But then another application of the inductive step tells us that the third object has the property as well:



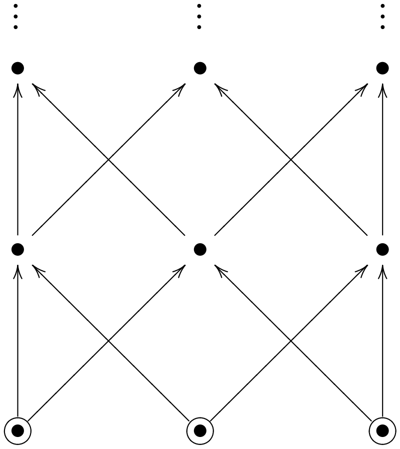
And so on; all objects in the line have the property:

⦿ ⦿ ⦿ ⦿ ···

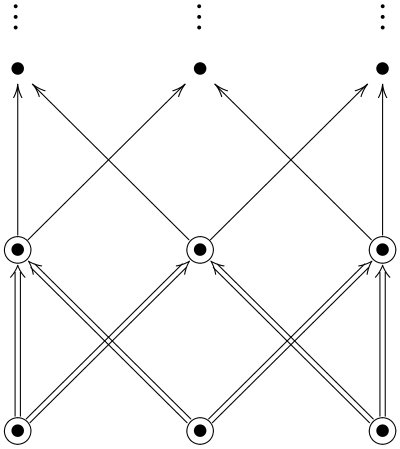
That is how induction works when applied to objects lined up in the manner depicted: there is a first object in line; after each object there is exactly one further object; and each object appears some finite number of jumps after the first object. Induction can also be applied to objects structured in different ways. Consider, for example, the following infinite grid of objects:



At the bottom of this grid there are three dots. Every pair of these three dots combines to produce one new dot. (For example, the leftmost dot on the second from the bottom level is produced by the leftmost two dots on the bottom level.) The resulting three dots (formed from the three pairs drawn from the three dots on the bottom level) form the second level of the grid. These three dots on the second level produce the third level in the same way, and so on. Suppose, now, that one could prove that the bottom three dots have some property:



(This is the "base case".) And suppose further that one could prove that whenever two dots with the property combine, the resulting dot also has the property ("inductive step"). Then, just as in the previous example, induction allows us to conclude that all the dots in the grid have the property. Given the base case and the inductive step, we know that the dots on the second level of the grid have the property:



But then, given the inductive step, we know that the dots on the third level have the property. And so on, for all the other levels.

In general, induction is a method for proving that each member of a certain collection of objects has a property. It works when (but only when) each object in the collection results from some "starting objects" by a finite number of iterations of some "operations". In the base case one proves that the starting objects have the property; in the induction step one proves that the operations *preserve* the property, in the sense that whenever one of the operations is applied to some objects with the property, the resulting new object has the property as well; and finally one concludes that all objects have the property.

This idea manifests itself in logic in a few ways. One is in a style of proof sometimes called "induction on formula construction" (or: induction "on the number of connectives" of the formula). Suppose we want to establish that absolutely every wff has a certain property, p. The method of proof by induction on formula construction tells us to first establish the following two claims:

b) every atomic wff (i.e. every sentence letter) has property p

i) for any wffs ϕ and ψ, *if* both ϕ and ψ have property p, *then* the wffs $\sim\phi$ and $\phi\rightarrow\psi$ also have property p

Once these are established, proof by induction allows us to conclude that every wff has property p. Why is this conclusion justified? Recall the definition of a wff from section 2.1: each wff is built up from atomic wffs by repeated application of clause ii): "if ϕ and ψ are wffs then $\sim\phi$ and $\phi\rightarrow\psi$ are also wffs". So each wff is the culmination of a finite process that starts with atomic wffs and continues by building conditionals and negations from wffs formed in previous steps of the process. But claim b) (the base case) shows that the *starting points* of this process all have property p. And claim i) (the induction step) shows that the subsequent steps in this process *preserve* property p: if the formulas one has built up so far have property p, then the next formula in the process (built up of previous formulas using either → or ∼) is guaranteed to also have p. So all wffs have property p. In terms of the general idea of inductive proof, the atomic wffs are our "starting objects" (like the bottom three dots in the grid), and the rules of grammar for ∼ and → which generate complex wffs from simpler wffs are the "operations".

Here is a simple example of proof by induction on formula construction:

Proof that every wff contains a finite number of sentence letters. We are trying to prove a statement of the form: every wff has property p. The property p in this case is *having a finite number of different sentence letters*. Our proof has two separate steps:

base case: here we must show that every atomic sentence has the property. This is obvious—atomic sentences are just sentence letters, and each of them contains one sentence letter, and thus finitely many different sentence letters.

induction step: here we must show that *if* wffs ϕ and ψ have property p, then so do $\sim\phi$ and $\phi\rightarrow\psi$. So we begin by *assuming*:

$$\text{formulas } \phi \text{ and } \psi \text{ each have finitely many different sentence letters} \qquad \text{(ih)}$$

This assumption is often called the "inductive hypothesis". And we must go on to show that both $\sim\phi$ and $\phi\rightarrow\psi$ have finitely many different sentence letters. This, too, is easy. $\sim\phi$ has as many different sentence letters as does ϕ; since ih) tells us that ϕ has finitely many, then so does $\sim\phi$. As for $\phi\rightarrow\psi$, it has, at most, $n+m$ sentence letters, where n and m are the number of different sentence letters in ϕ and ψ, respectively; ih) tells us that n and m are finite, and so $n+m$ is finite as well.

We've shown that every atomic formula has the property *having a finite number of different sentence letters*; and we've shown that the property is inherited by complex formulas built according to the formation rules. But every wff is either atomic, or built from atomics by a finite series of applications of the formation rules. Therefore, by induction, every wff has the property. ■

A different form of inductive proof is called for in the following proof of soundness:

Proof of soundness for PL. Unlike the previous inductive proof, here we are not trying to prove something of the form "Every wff has property p". Instead, we're trying to prove something of the form "Every *theorem* has property p". Nevertheless we can still use induction, only we need to use induction of a slightly different sort from induction on formula construction. Consider: a theorem is any line of a proof. And every line of every proof is the culmination of a finite series of wffs, in which each member is either an axiom, or follows from earlier lines by modus ponens. So the conditions are right for an inductive proof. The "starting points" are the axioms; and the "operation" is the inference of a new line from earlier lines using modus ponens. If we can show that the starting points (axioms) have the property of validity, and that the operation (modus ponens) preserves the property of validity, then we can conclude that every wff in every proof—i.e., every theorem—has the property of validity. This sort of inductive proof is called induction "on the proof of a formula" (or induction "on the length of the formula's proof").

base case: here we need to show that every PL-axiom is valid. This is tedious but straightforward. Take PL1, for example. Suppose for reductio that some instance of PL1 is invalid, i.e., for some PL-interpretation $\mathscr{I}$,

$V_{\mathscr{I}}(\phi\rightarrow(\psi\rightarrow\phi)) = 0$. Thus, $V_{\mathscr{I}}(\phi) = 1$ and $V_{\mathscr{I}}(\psi\rightarrow\phi) = 0$. Given the latter, $V_{\mathscr{I}}(\phi) = 0$—contradiction. Analogous proofs can be given that instances of PL2 and PL3 are also valid (exercise 2.5).

induction step: here we begin by *assuming* that every line in a proof up to a certain point is valid (the "inductive hypothesis"); we then show that if one adds another line that follows from earlier lines by the rule modus ponens, that line must be valid too. I.e., we're trying to show that "modus ponens preserves validity". So, assume the inductive hypothesis: that all the earlier lines in the proof are valid. And now, consider the result of applying modus ponens. That means that the new line we've added to the proof is some formula ψ, which we've inferred from two earlier lines that have the forms $\phi\rightarrow\psi$ and ϕ. We must show that ψ is a valid formula, i.e., is true in every interpretation. So let $\mathscr{I}$ be any interpretation. By the inductive hypothesis, all earlier lines in the proof are valid, and hence both $\phi\rightarrow\psi$ and ϕ are valid. Thus, $V_{\mathscr{I}}(\phi) = 1$ and $V_{\mathscr{I}}(\phi\rightarrow\psi) = 1$. But if $V_{\mathscr{I}}(\phi) = 1$ then $V_{\mathscr{I}}(\psi)$ can't be 0, for if it were, then $V_{\mathscr{I}}(\phi\rightarrow\psi)$ would be 0, and it isn't. Thus, $V_{\mathscr{I}}(\psi) = 1$.

(If our system had included rules other than modus ponens, we would have needed to show that they too preserve validity. The paucity of rules in axiomatic systems makes the construction of proofs within those systems a real pain in the neck, but now we see how it makes metalogical life easier.)

We've shown that the axioms are valid, and that modus ponens preserves validity. All theorems are generated from the axioms via modus ponens in a finite series of steps. So, by induction, every theorem is valid. ■

One nice thing about soundness is that it lets us establish facts of *unprovability*. Soundness says "if $\vdash \phi$ then $\vDash \phi$". Equivalently, it says: "if $\nvDash \phi$ then $\nvdash \phi$". So, to show that something isn't a theorem, it suffices to show that it isn't valid. Consider, for example, the formula $(P\rightarrow Q)\rightarrow(Q\rightarrow P)$. There exist PL-interpretations in which the formula is false, namely, PL-interpretations in which P is 0 and Q is 1. So, $(P\rightarrow Q)\rightarrow(Q\rightarrow P)$ is not valid (since it's not true in all PL-interpretations.) But then soundness tells us that it isn't a theorem either. In general: given soundness, in order to show that a formula isn't a theorem, all you need to do is find an interpretation in which it isn't true.

Before we leave this section, let me reiterate the distinction between the two types of induction most commonly used in metalogic. Induction on the proof of a formula (the type of induction used to establish soundness) is used when one is establishing a fact of the form: *every **theorem** has a certain property p*. Here the base case consists of showing that the axioms have the property p, and the inductive step consists of showing that the rules of inference preserve p—i.e., in the case of modus ponens: that *if* ϕ and $\phi\rightarrow\psi$ both have property p then so does ψ. (Induction on proofs can also be used to show that all wffs *provable from* a given set Γ have a given property; in that case the base case would also need to include a demonstration that all members of Γ have the property.) Induction on formula construction (the type of induction used to show that all formulas have finitely many sentence letters), on the other hand, is used when one is trying to establish a fact of the form: *every* ***formula*** *has a certain property p*. Here the base case consists of showing that all sentence letters have property p; and the inductive step consists of showing that the rules of formation preserve p—i.e., that *if* ϕ and ψ both have property p, then both $(\phi\rightarrow\psi)$ and $\sim\phi$ also will have property p.

If you're ever proving something by induction, it's important to identify what sort of inductive proof you're constructing. What are the entities you're dealing with? What is the property p? What are the starting points, and what are the operations generating new entities from the starting points? If you're trying to construct an inductive proof and get stuck, you should return to these questions and make sure you're clear about their answers.

Exercise 2.5 Finish the soundness proof by showing that all instances of axiom schemas PL2 and PL3 are valid.

Exercise 2.6 Consider the following (strange) system of propositional logic. The definition of wffs is the same as for standard propositional logic, and the rules of inference are the same (just one rule: modus ponens); but the axioms are different. For any wffs ϕ and ψ, the following are axioms:

$$\phi \to \phi$$
$$(\phi \to \psi) \to (\psi \to \phi)$$

Establish the following two facts about this system: (a) every theorem of this system has an even number of "∼"s; (b) soundness is false for this system—i.e., some theorems are not valid formulas.

Exercise 2.7 Show by induction that the truth value of a wff depends only on the truth values of its sentence letters. That is, show that for any wff ϕ and any PL-interpretations $\mathscr{I}$ and $\mathscr{I}'$, if $\mathscr{I}(\alpha) = \mathscr{I}'(\alpha)$ for each sentence letter α in ϕ, then $V_{\mathscr{I}}(\phi) = V_{\mathscr{I}'}(\phi)$.

Exercise 2.8** Suppose that a wff ϕ has no repetitions of sentence letters (i.e., each sentence letter occurs at most once in ϕ.) Show that ϕ is not PL-valid.

Exercise 2.9 Prove "strong soundness": for any set of formulas, Γ, and any formula ϕ, if $\Gamma \vdash \phi$ then $\Gamma \vDash \phi$ (i.e., if ϕ is provable from Γ then ϕ is a semantic consequence of Γ.)

Exercise 2.10** Prove the soundness of the sequent calculus. That is, show that if $\Gamma \Rightarrow \phi$ is a provable sequent, then $\Gamma \vDash \phi$. (No need to go through each and every detail of the proof once it becomes repetitive.)

2.9 Completeness of PL

We're finally ready for the completeness proof. We will give what is known as a "Henkin-proof", after Leon Henkin, who used similar methods to demonstrate completeness for (nonmodal) predicate logic. Most of the proof will consist of assembling various pieces—various definitions and facts. The point of these pieces will become apparent at the end, when we put them all together.

2.9.1 Maximal consistent sets of wffs

Let "$\bot$" abbreviate "$\sim(P\rightarrow P)$". (The idea of $\bot$ is that it stands for a generic contradiction. The choice of $\sim(P\rightarrow P)$ was arbitrary; all that matters is that $\bot$ is the negation of a theorem.) Here are the central definitions we'll need:

Definition of consistency and maximality:

- A set of wffs, Γ, is inconsistent iff $\Gamma \vdash \bot$. Γ is consistent iff it is not inconsistent
- A set of wffs, Γ, is maximal iff for every wff ϕ, either ϕ or $\sim\phi$ (or perhaps both) is a member of Γ

Intuitively: a maximal set is so large that it contains each formula or its negation; and a consistent set is one from which you can't prove a contradiction. Note the following lemmas:

Lemma 2.1 For any set of wffs Γ and wff ϕ, if ϕ is provable from Γ then ϕ is provable from some finite subset of Γ. That is, if $\Gamma \vdash \phi$ then $\gamma_1 \dots \gamma_n \vdash \phi$ for some $\gamma_1 \dots \gamma_n \in \Gamma$ (or else $\vdash \phi$)

Proof. If $\Gamma \vdash \phi$ then there is some proof, A, of ϕ from Γ. Like every proof, A is a finite series of wffs. Thus, only finitely many of Γ's members can have occurred as lines in A. Let $\gamma_1 \dots \gamma_n$ be those members of Γ. (If no member of Γ occurs in A then A proves ϕ from no premises at all, in which case $\vdash \phi$.) In addition to counting as a proof of ϕ from Γ, proof A is also a proof of ϕ from $\{\gamma_1 \dots \gamma_n\}$. Thus, $\gamma_1 \dots \gamma_n \vdash \phi$. ■

Lemma 2.2 For any set of wffs Γ, if $\Gamma \vdash \phi$ and $\Gamma \vdash \sim\phi$ for some ϕ then Γ is inconsistent

Proof. Follows immediately from ex falso quodlibet (example 2.11) and Cut. ■

Note that the first lemma tells us that a set is inconsistent iff some finite subset of that set is inconsistent.

2.9.2 Maximal consistent extensions

Suppose we begin with a consistent set Δ that isn't maximal—for at least one wff ϕ, Δ contains neither ϕ nor $\sim\phi$. Is there some way of adding wffs to Δ to make it maximal, without destroying its consistency? That is, is Δ guaranteed to have some maximal consistent "extension"? The following theorem tells us that the answer is *yes*:

Theorem 2.3 If Δ is a consistent set of wffs, then there exists some maximal consistent set of wffs, Γ, such that $\Delta \subseteq \Gamma$

Proof of Theorem 2.3. In outline, we're going to build up Γ as follows. We're going to start by dumping all the formulas in Δ into Γ. Then we will go through *all* the wffs, ϕ_1, ϕ_2,..., one at a time. For each wff, we're going to dump either it or its negation into Γ, depending on which choice would be consistent. After we're done, our set Γ will obviously be maximal; it will obviously contain Δ as a subset; and, we'll show, it will also be consistent.

So, let ϕ_1, ϕ_2,... be a list—an infinite list, of course—of all the wffs.[10] To

[10] We need to be sure that there is some way of arranging all the wffs into such a list. Here is one method. First, begin with a list of the primitive expressions of the language. In the case of PL this can be done as follows:

$$\begin{array}{ccccccc} (&) & \sim & \rightarrow & P_1 & P_2 & \dots \\ 1 & 2 & 3 & 4 & 5 & 6 & \dots \end{array}$$

(For simplicity, get rid of all the sentence letters except for $P_1, P_2, \dots$.) Since we'll need to refer to what *position* an expression has in this list, the positions of the expressions are listed underneath those expressions. (E.g., the position of the → is 4.) Now, where ϕ is any wff, call the *rating* of ϕ the sum of the positions of the occurrences of its primitive expressions. (The rating for the wff $(P_1 \rightarrow P_1)$, for example, is $1+5+4+5+2=17$.) We can now construct the listing of all the wffs of PL by an infinite series of stages: stage 1, stage 2, etc. In stage n, we append to our growing list all the wffs of rating n, *in alphabetical order*. The notion of alphabetical order here is the usual one, given the ordering of the primitive expressions laid out above. (E.g., just as 'and' comes before 'dna' in alphabetical order, since 'a' precedes 'd' in the usual ordering of the English alphabet, $(P_1 \rightarrow P_2)$ comes before $(P_2 \rightarrow P_1)$ in alphabetical order since P_1 comes before P_2 in the ordering of the alphabet of PL. Note that each of these wffs are inserted into the list in stage 18, since each has rating 18.) In stages 1–4 no wffs are added at all, since every wff must have at least one sentence letter and P_1 is the sentence letter with the smallest position. In stage 5 there is one wff: P_1. Thus, the first member of our list of wffs is P_1. In stage 6 there is one wff: P_2, so P_2 is the second member of the list. In every subsequent stage there are only finitely many wffs; so each stage adds finitely many wffs to the list; each wff gets added at some stage; so each wff eventually gets added after some finite amount of time to this list.

construct Γ, our strategy is to start with Δ, and then go through this list one-by-one, at each point adding either ϕ_i or $\sim\phi_i$. Here's how we do this more carefully. We first define an infinite sequence of sets, $\Gamma_0, \Gamma_1, \dots$:

$$\Gamma_0 = \Delta$$
$$\Gamma_{n+1} = \begin{cases} \Gamma_n \cup \{\phi_{n+1}\} & \text{if } \Gamma_n \cup \{\phi_{n+1}\} \text{ is consistent} \\ \Gamma_n \cup \{\sim\phi_{n+1}\} & \text{if } \Gamma_n \cup \{\phi_{n+1}\} \text{ is not consistent} \end{cases}$$

This definition is recursive, notice. We begin with a noncircular definition of the first member of the sequence of sets, Γ_0, and after that, we define each subsequent member Γ_{n+1} in terms of the previous member Γ_n: we add ϕ_{n+1} to Γ_n if the result of doing so would be consistent; otherwise we add $\sim\phi_{n+1}$.

Next let's prove that each member in this sequence—that is, each Γ_i—is a consistent set. We do this inductively, by first showing that Γ_0 is consistent, and then showing that if Γ_n is consistent, then so will be Γ_{n+1}. This is a different sort of inductive proof from what we've seen so far, neither an induction on formula construction nor on formula proof. Nevertheless we have the required structure for proof by induction: each of the objects of interest (the Γ_is) is generated from a starting point (Γ_0) by a finite series of operations (the operation taking us from Γ_n to Γ_{n+1}).

Base case: obviously, Γ_0 is consistent, since Δ was stipulated to be consistent.

Inductive step: we suppose that Γ_n is consistent (inductive hypothesis), and then show that Γ_{n+1} is consistent. Look at the definition of Γ_{n+1}. What Γ_{n+1} gets defined as depends on whether $\Gamma_n \cup \{\phi_{n+1}\}$ is consistent. If $\Gamma_n \cup \{\phi_{n+1}\}$ *is* consistent, then Γ_{n+1} gets defined as that very set $\Gamma_n \cup \{\phi_{n+1}\}$. So of course Γ_{n+1} is consistent in that case.

The remaining possibility is that $\Gamma_n \cup \{\phi_{n+1}\}$ is *in*consistent. In that case, Γ_{n+1} gets defined as $\Gamma_n \cup \{\sim\phi_{n+1}\}$. So must show that in this case, $\Gamma_n \cup \{\sim\phi_{n+1}\}$ is consistent. Suppose for reductio that it isn't. Then ⊥ is provable from $\Gamma_n \cup \{\sim\phi_{n+1}\}$, and so given lemma 2.1 is provable from some finite subset of this set; and the finite subset must contain $\sim\phi_{n+1}$ since Γ_n was consistent. Letting $\psi_1 \dots \psi_m$ be the remaining members of the finite subset, we have, then: $\psi_1 \dots \psi_m, \sim\phi_{n+1} \vdash \bot$, from which we get $\psi_1 \dots \psi_m \vdash \sim\phi_{n+1} \rightarrow \bot$ by the deduction theorem. Since $\Gamma_n \cup \{\phi_{n+1}\}$ is inconsistent, similar reasoning tells us that $\chi_1 \dots \chi_p \vdash \phi_{n+1} \rightarrow \bot$, for some $\chi_1 \dots \chi_p \in \Gamma_n$. It then follows by "excluded middle MP" (example 2.11) and Cut that $\psi_1 \dots \psi_m, \chi_1 \dots \chi_p \vdash \bot$. Since $\psi_1 \dots \psi_m, \chi_1 \dots \chi_p$ are all members of Γ_n, this contradicts the fact that Γ_n is consistent.

We have shown that all the sets in our sequence Γ_i are consistent. Let us now define Γ to be the *union* of all the sets in the infinite sequence—i.e., $\{\phi : \phi \in \Gamma_i$ for some $i\}$. We must now show that Γ is the set we're after: that i) $\Delta \subseteq \Gamma$, ii) Γ is maximal, and iii) Γ is consistent.

Any member of Δ is a member of Γ_0 (since Γ_0 was defined as Δ), hence is a member of one of the Γ_is, and hence is a member of Γ. So $\Delta \subseteq \Gamma$.

Any wff is in the list of all the wffs somewhere—i.e., it is ϕ_i for some i. But by definition of Γ_i, either ϕ_i or $\sim\phi_i$ is a member of Γ_i; and so one of these is a member of Γ. Γ is therefore maximal.

Suppose for reductio that Γ is inconsistent. Given lemma 2.1, there exist $\psi_1 \dots \psi_m \in \Gamma$ such that $\psi_1 \dots \psi_m \vdash \bot$. By definition of Γ, each $\psi_i \in \Gamma_{j_i}$, for some j_i. Let k be the largest of $j_1 \dots j_m$. Given the way the $\Gamma_0, \Gamma_1, \dots$ series is constructed, each set in the series is a subset of all subsequent ones. Thus, each of $\psi_1 \dots \psi_m$ is a member of Γ_k, and thus Γ_k is inconsistent. But we showed that each member of the series $\Gamma_0, \Gamma_1, \dots$ is consistent. ■

2.9.3 Features of maximal consistent sets

Next we'll establish two facts about maximal consistent sets that we'll need for the completeness proof:

Lemma 2.4 Where Γ is any maximal consistent set of wffs:

2.4a for any wff ϕ, exactly one of ϕ, $\sim\phi$ is a member of Γ

2.4b $\phi\rightarrow\psi \in \Gamma$ iff either $\phi \notin \Gamma$ or $\psi \in \Gamma$

Proof of Lemma 2.4a. Since Γ is maximal it must contain at least one of ϕ or $\sim\phi$. But it cannot contain both; otherwise each would be provable from Γ, whence by lemma 2.2, Γ would be inconsistent. ■

Proof of Lemma 2.4b. Suppose first that $\phi\rightarrow\psi$ is in Γ, and suppose for reductio that ϕ is in Γ but ψ is not. Then we can prove ψ from Γ (begin with ϕ and $\phi\rightarrow\psi$ as premises, and then use MP). But since $\psi \notin \Gamma$ and Γ is maximal, $\sim\psi$ is in Γ, and hence is provable from Γ. Given lemma 2.2, this contradicts Γ's consistency.

Suppose for the other direction that either $\phi \notin \Gamma$ or $\psi \in \Gamma$, and suppose for reductio that $\phi\rightarrow\psi \notin \Gamma$. Since Γ is maximal, $\sim(\phi\rightarrow\psi) \in \Gamma$. Then $\Gamma \vdash \sim(\phi\rightarrow\psi)$, and so by "negated conditional" (example 2.11) and Cut, $\Gamma \vdash \phi$ and $\Gamma \vdash \sim\psi$.

Now, if $\phi \notin \Gamma$ then $\sim\phi \in \Gamma$ and so $\Gamma \vdash \sim\phi$; and if on the other hand $\psi \in \Gamma$ then $\Gamma \vdash \psi$. Each possibility contradicts Γ's consistency, given lemma 2.2. ■

2.9.4 The proof

Now it's time to put together all the pieces that we've assembled.

Proof of PL completeness. Completeness says: if $\vDash \phi$ then $\vdash \phi$. We'll prove this by proving the equivalent statement: if $\nvdash \phi$ then $\nvDash \phi$. So, suppose that $\nvdash \phi$. We must construct some PL-interpretation in which ϕ isn't true.

Since $\nvdash \phi$, $\{\sim\phi\}$ must be consistent. For suppose otherwise. Then $\sim\phi \vdash \bot$; so $\vdash \sim\phi \to \bot$ by the deduction theorem. That is, given the definition of $\bot$: $\vdash \sim\phi \to \sim(P \to P)$. Then by contraposition 1 (example 2.11), $\vdash (P \to P) \to \phi$. But $\vdash P \to P$ (exercise 2.4a), and so $\vdash \phi$—contradiction.

Since $\{\sim\phi\}$ is consistent, theorem 2.3 tells us that it is a subset of some maximal consistent set of wffs Γ. Next, let's use Γ to construct a somewhat odd PL-interpretation. This PL-interpretation decides whether a sentence letter is true or false by looking to see whether that sentence letter *is a member of* Γ. What we will do next is show that *all* formulas, not just sentence letters, are true in this odd interpretation iff they are members of Γ.

So, let $\mathscr{I}$ be the PL-interpretation in which for any sentence letter α, $\mathscr{I}(\alpha) = 1$ iff $\alpha \in \Gamma$. We must show that:

$$\text{for every wff } \phi,\ V_{\mathscr{I}}(\phi) = 1 \text{ iff } \phi \in \Gamma \qquad (*)$$

We do this by induction on formula construction. The base case, that the assertion holds for sentence letters, follows immediately from the definition of $\mathscr{I}$. Next we make the inductive hypothesis (ih): that wffs ϕ and ψ are true in $\mathscr{I}$ iff they are members of Γ, and we show that the same is true of $\sim\phi$ and $\phi \to \psi$.

First, $\sim\phi$: we must show that $V_{\mathscr{I}}(\sim\phi) = 1$ iff $\sim\phi \in \Gamma$:[11]

$$\begin{aligned} V_{\mathscr{I}}(\sim\phi) = 1 &\text{ iff } V_{\mathscr{I}}(\phi) = 0 && \text{(truth cond. for} \sim\text{)}\\ &\text{ iff } \phi \notin \Gamma && \text{(ih)}\\ &\text{ iff } \sim\phi \in \Gamma && \text{(lemma 2.4a)} \end{aligned}$$

[11] Here we continue to use the fact that a formula has one truth value iff it lacks the other.

Next, →: we must show that $V_{\mathscr{I}}(\phi\rightarrow\psi)=1$ iff $\phi\rightarrow\psi\in\Gamma$:

$$\begin{aligned}
V_{\mathscr{I}}(\phi\rightarrow\psi)=1 &\text{ iff either } V_{\mathscr{I}}(\phi)=0 \text{ or } V_{\mathscr{I}}(\psi)=1 && \text{(truth cond for }\rightarrow\text{)}\\
&\text{ iff either } \phi\notin\Gamma \text{ or } \psi\in\Gamma && \text{(ih)}\\
&\text{ iff } \phi\rightarrow\psi\in\Gamma && \text{(lemma 2.4b)}
\end{aligned}$$

The inductive proof of (*) is complete. But now, since $\{\sim\phi\}\subseteq\Gamma$, $\sim\phi\in\Gamma$, and so by lemma 2.4a, $\phi\notin\Gamma$. Thus, by (*), ϕ is not true in $\mathscr{I}$. So we have succeeded in constructing an interpretation in which ϕ isn't true.

■

Chapter 6

Propositional Modal Logic

Modal logic is the logic of necessity and possibility. In it we treat "modal" words like 'necessary', 'possible', 'can', and 'must' as logical constants. Our new symbols for these words are called "modal operators":

$\Box\phi$: "It is necessary that ϕ" (or: "Necessarily, ϕ", "It must be that ϕ")

$\Diamond\phi$: "It is possible that ϕ" (or: "Possibly, ϕ", "It could be that ϕ", "It can be that ϕ", "It might be that ϕ", "it might have been that ϕ")

It helps to think of modality in terms of *possible worlds*. A possible world is a *complete* and *possible* scenario. Calling a scenario "possible" means simply that it's possible in the broadest sense for the scenario to happen. This requirement disqualifies scenarios in which, for example, it is both raining and also not raining (at the same time and place)—such a thing couldn't happen, and so doesn't happen in any possible world. But within this limit, we can imagine all sorts of possible worlds: possible worlds with talking donkeys, possible worlds in which I am ten feet tall, and so on. "Complete" means simply that no detail is left out—possible worlds are completely *specific* scenarios. There is no possible world in which I am "somewhere between ten and eleven feet tall" without being some particular height.[1] Likewise, in any possible world in which I am exactly ten feet, six inches tall (say), I must have some particular weight, must live in some particular place, and so on. One of these possible worlds is the actual world—this is the complete and possible scenario that in fact obtains.

[1] This is not to say that possible worlds exclude vagueness.

The rest of them are merely possible—they do not obtain, but would have obtained if things had gone differently. In terms of possible worlds, we can think of our modal operators thus:

"$\Box\phi$" is true iff ϕ is true in *all* possible worlds

"$\Diamond\phi$" is true iff ϕ is true in *at least one* possible world

It is necessarily true that all bachelors are male; in every possible world, every bachelor is male. There might have existed a talking donkey; some possible world contains a talking donkey.

Possible worlds provide, at the very least, a vivid way to think about necessity and possibility. How much more they provide is an open philosophical question. Some maintain that possible worlds are the key to the metaphysics of modality, that *what it is* for a proposition to be necessarily true is for it to be true in all possible worlds.[2] Whether this view is defensible is a question beyond the scope of this book; what is important for present purposes is that we distinguish possible worlds as a vivid heuristic from possible worlds as a concern in serious metaphysics.

Natural language modal words are semantically flexible in a systematic way. For example, suppose I say that I can't attend a certain conference in Cleveland. What is the force of "can't" here? Probably I'm saying that my attending the conference is inconsistent with honoring other commitments I've made at that time. But notice that another sentence I might utter is: "I *could* attend the conference; but I would have to cancel my class, and I don't want to do that." Now I've said that I *can* attend the conference; have I contradicted my earlier assertion that I cannot attend the conference? No—what I mean now is perhaps that I have the means to get to Cleveland on that date. I have shifted what I mean by "can".

In fact, there is quite a wide range of things one can mean by words for possibility:

I can come to the party, but I can't stay late. ("can" = "is not inconvenient")

Humans can travel to the moon, but not Mars. ("can" = "is achievable with current technology")

[2] Sider (2003) presents an overview of this topic.

> *It's possible to move almost as fast as the speed of light, but not to travel faster than light.* ("possible" = "is consistent with the laws of nature")
>
> *Objects could have traveled faster than the speed of light (if the laws of nature had been different), but no matter what the laws had been, nothing could have traveled faster than itself.* ("could" = "metaphysical possibility")
>
> *You may borrow but you may not steal.* ("may" = "morally acceptable")
>
> *It might rain tomorrow* ("might" = "epistemic possibility")

For any strength of possibility, there is a corresponding strength of necessity, since "necessarily ϕ" is equivalent to "not-possibly-not-ϕ". (Similarly, "possibly ϕ" is equivalent to "not-necessarily-not-ϕ".) So we have a range of strengths of necessity as well: natural necessity (guaranteed by the laws of nature), moral or "deontic" necessity (required by morality), epistemic necessity ("known to be true") and so on.

Some sorts of necessity imply truth; those that do are called "alethic" necessities. For example, if P is known then P is true; if it is naturally necessary that massive particles attract one another, then massive particles do in fact attract one another. Epistemic and natural necessity are alethic. Deontic necessity, on the other hand, is not alethic; we do not always do what is morally required.

As we saw, we can think of the $\Box$ and the $\Diamond$ as quantifiers over possible worlds (the former a universal quantifier, the latter an existential quantifier). This idea can accommodate the fact that necessity and possibility come in different strengths: those different strengths result from different restrictions on the quantifiers over possible worlds. Thus, natural possibility is truth in some possible world that obeys the actual world's laws; deontic possibility is truth in some possible world in which nothing morally forbidden occurs; and so on.[3]

[3] This raises a question, though: to what strength of 'necessary' and 'possible' does the notion of possible world itself correspond? Is there some special, strictest notion of necessity, which can be thought of as truth in absolutely all possible worlds? Or do we simply have different notions of possible world corresponding to different strengths of necessity?

6.1 Grammar of MPL

Our first topic in modal logic is the addition of the $\Box$ and the $\Diamond$ to propositional logic; the result is *modal propositional logic* ("MPL"). A further step will be modal predicate logic (chapter 9).

We need a new language: the language of MPL. The grammar of this language is just like the grammar of propositional logic, except that we add the $\Box$ as a new one-place sentence connective:

Primitive vocabulary:

- Sentence letters: $P, Q, R\dots$, with or without numerical subscripts
- Connectives: $\rightarrow$, $\sim$, $\Box$
- Parentheses: (,)

Definition of wff:

- Sentence letters are wffs
- If ϕ and ψ are wffs then $(\phi\rightarrow\psi)$, $\sim\phi$, and $\Box\phi$ are also wffs
- Only strings that can be shown to be wffs using the preceding clauses are wffs

The $\Box$ is the only new primitive connective. But just as we were able to define $\wedge$, $\vee$, and $\leftrightarrow$, we can define new nonprimitive modal connectives:

- "$\Diamond\phi$" ("Possibly ϕ") is short for "$\sim\Box\sim\phi$"
- "ϕ⥽ψ" ("ϕ strictly implies ψ") is short for "$\Box(\phi\rightarrow\psi)$"

6.2 Symbolizations in MPL

Modal logic allows us to symbolize a number of sentences we couldn't symbolize before. The most obvious cases are sentences that overtly involve "necessarily", "possibly", or equivalent expressions:

> Necessarily, if snow is white, then snow is white or grass is green
>
> $\Box[S\rightarrow(S\vee G)]$

I'll go if I must
$\Box G \to G$

It is possible that Bush will lose the election
$\Diamond L$

Snow might have been either green or blue
$\Diamond(G \vee B)$

If snow could have been green, then grass could have been white
$\Diamond G \to \Diamond W$

'Impossible' and related expressions signify the lack of possibility:

It is impossible for snow to be both white and not white
$\sim\Diamond(W \wedge \sim W)$

If grass cannot be clever then snow cannot be furry
$\sim\Diamond C \to \sim\Diamond F$

God's being merciful is inconsistent with your imperfection being incompatible with your going to heaven
$\sim\Diamond(M \wedge \sim\Diamond(I \wedge H))$

As for the strict conditional, it arguably does a decent job of representing certain English conditional constructions:

Snow is a necessary condition for skiing
$\sim W \rightfishtail \sim K$

Food and water are required for survival
$\sim(F \wedge W) \rightfishtail \sim S$

Thunder implies lightning
$T \rightfishtail L$

Once we add modal operators, we can make an important distinction involving modal conditionals in natural language. Consider the sentence "if Jones is a bachelor, then he must be unmarried". The surface grammar misleadingly suggests the symbolization:

$$B \to \Box U$$

But suppose that Jones is in fact a bachelor. It would then follow from this symbolization that the proposition that Jones is unmarried is necessarily true. But nothing we have said suggests that Jones is *necessarily* a bachelor. Surely Jones *could* have been married! In fact, one would normally *not* use the sentence "if Jones is a bachelor, then he must be unmarried" to mean that if Jones is in fact a bachelor, then the following is a necessary truth: Jones is unmarried. Rather, one would mean: necessarily, if Jones is a bachelor then Jones is unmarried:

$$\Box(B\to U)$$

It is the *relationship* between Jones's being a bachelor and his being unmarried that is necessary. Think of this in terms of possible worlds: the first symbolization says that if Jones is a bachelor in the actual world, then Jones is unmarried in every possible world (which is absurd); whereas the second one says that in each possible world, w, if Jones is a bachelor *in* w, then Jones is unmarried *in* w (which is quite sensible). The distinction between $\phi\to\Box\psi$ and $\Box(\phi\to\psi)$ is called the distinction between the "necessity of the consequent" (first sentence) and the "necessity of the consequence" (second sentence). It is important to keep the distinction in mind, because of the fact that English surface structure is misleading.

One final point: when representing English sentences using the $\Box$ and the $\Diamond$, keep in mind that these expressions can be used to express different strengths of necessity and possibility. (One could introduce different symbols for the different sorts; we'll do a bit of this in chapter 7.)

6.3 Semantics for MPL

As usual, we'll consider semantics first. We'll show how to construct mathematical configurations in a way that's appropriate to modal logic, and show how to define truth for formulas of MPL within these configurations. Ideally, we'd like the assignment of truth values to wffs to mirror the way that natural language modal statements are made true by the real world, so that we can shed light on the meanings of natural language modal words, and in order to provide plausible semantic models of the notions of logical truth and logical consequence.

In constructing a semantics for MPL, we face two main challenges, one philosophical, the other technical. The philosophical challenge is simply that it isn't wholly clear which formulas of MPL are indeed logical truths. It's hard to construct an engine to spit out logical truths if you don't know which logical truths you want it to spit out. With a few exceptions, there is widespread agreement over which formulas of nonmodal propositional and predicate logic are logical truths. But for modal logic this is less clear, especially for sentences that contain iterations of modal operators. Is $\Box P \to \Box\Box P$ a logical truth? It's hard to say.

A quick peek at the history of modal logic is in order. Modal logic arose from dissatisfaction with the material conditional → of standard propositional logic. In standard logic, $\phi \to \psi$ is true whenever ϕ is false or ψ is true; but in expressing the conditionality of ψ on ϕ, we sometimes want to require a tighter relationship: we want it not to be a mere *accident* that either ϕ is false or ψ is true. To express this tighter relationship, C. I. Lewis introduced the strict conditional $\phi \strictif \psi$, which he defined, as above, as $\Box(\phi \to \psi)$.[4] Thus defined, $\phi \strictif \psi$ isn't automatically true just because ϕ is false or ψ is true. It must be *necessarily true* that either ϕ is false or ψ is true.

Lewis then asked: what principles govern this new symbol □? Certain principles seemed clearly appropriate, for instance: $\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$. Others were less clear. Is $\Box\phi \to \Box\Box\phi$ a logical truth? What about $\Diamond\Box\phi \to \phi$?

Lewis's solution to this problem was not to choose. Instead, he formulated several different *modal systems*. He did this axiomatically, by formulating different systems that differed from one another by containing different axioms and hence different theorems.

We will follow Lewis's approach, and construct several different modal systems. Unlike Lewis, we'll do this semantically at first (the semantics for modal logic we will study was published by Saul Kripke in the 1950s, long after Lewis was writing), by constructing different definitions of a model for modal logic. The definitions will differ from one another in ways that result in different sets of valid formulas. In section 6.4 we'll study Lewis's axiomatic systems, and in sections 6.5 and 6.6 we'll discuss the relationship between the semantics and the axiom systems.

Formulating multiple systems does not answer the philosophical question of which formulas of modal logic are logically true; it merely postpones it. The question re-arises when we want to *apply* Lewis's systems; when we ask which system is the *correct* system—i.e., which one correctly mirrors the logical properties of the *English* words 'possibly' and 'necessarily'? (Note that since there are different sorts of necessity and possibility, different systems might

[4] See Lewis (1918); Lewis and Langford (1932).

correctly represent different sorts.) But I'll mostly ignore such philosophical questions here.

The technical challenge to constructing a semantics for MPL is that the modal operators $\Box$ and $\Diamond$ are not truth functional. A sentential connective is truth-functional iff whenever it combines with sentences to form a new sentence, the truth value of the resulting sentence is determined by the truth values of the component sentences. For example, 'it is not the case that' is truth-functional because the truth value of "it is not the case that ϕ" is determined by the truth value of ϕ. But 'necessarily' is not truth-functional. If I tell you that ϕ is true, you won't yet have enough information to determine whether "Necessarily ϕ" is true or false, since you won't know whether ϕ is necessarily true or merely contingently true. Here's another way to put the point: even though the sentences "If Ted is a philosopher then Ted is a philosopher" and "Ted is a philosopher" have the same truth value, if you prefix each with 'Necessarily' (intended to mean metaphysical necessity, say), you get sentences with different truth values. Hence, the truth value of "Necessarily ϕ" is not a function of the truth value of ϕ. Similarly, 'possibly' isn't truth-functional either: 'I might have been six feet tall' is true, whereas 'I might have been a round square' is false, despite the sad fact that 'I am six feet tall' and 'I am a round square' have the same truth value.

Since the $\Box$ and the $\Diamond$ are supposed to represent 'necessarily' and 'possibly', and since the latter aren't truth-functional, we can't do modal semantics with truth tables. For the method of truth tables assumes truth-functionality. Truth tables are just pictures of truth functions: they specify what truth value a complex sentence has as a function of what truth values its parts have. Our challenge is clear: we need a semantics for the $\Box$ and the $\Diamond$ other than the method of truth tables.

6.3.1 Kripke models

Our approach will be that of *possible-worlds semantics*. The intuitive idea is to count $\Box\phi$ as being true iff ϕ is true in all possible worlds, and $\Diamond\phi$ as being true iff ϕ is true in some possible worlds. More carefully: we are going to develop models for modal propositional logic. These models will contain objects we will call "possible worlds". And formulas are going to be true or false "in" (or "at") these worlds. That is, we are going to assign truth values to formulas in these models relative to possible worlds, rather than absolutely. Truth values of propositional-logic compound formulas—that is, negations and conditionals—

will be determined by truth tables within each world; $\sim\phi$, for example, will be true at a world iff ϕ is false at that world. But the truth value of $\Box\phi$ at a world won't be determined by the truth value of ϕ at that world; the truth value of ϕ at *other* worlds will also be relevant.

Specifically, $\Box\phi$ will count as true at a world iff ϕ is true at every world that is "accessible" from the first world. What does "accessible" mean? Each model will come equipped with a binary relation, $\mathscr{R}$, over the set of possible worlds; we will say that world v is "accessible from" world w when $\mathscr{R}wv$. The intuitive idea is that $\mathscr{R}wv$ if and only if v is *possible relative to* w. That is, if you live in world w, then from your perspective, the events in world v are possible.

The idea that what is possible might vary depending on what possible world you live in might at first seem strange, but it isn't really. "It is physically impossible to travel faster than the speed of light" is true in the actual world, but false in worlds where the laws of nature allow faster-than-light travel.

On to the semantics. We first define a generic notion of an MPL model, which we'll then use to give a semantics for different modal systems:

Definition of model: An MPL-model is an ordered triple, $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$, where:

- $\mathscr{W}$ is a non-empty set of objects ("possible worlds")
- $\mathscr{R}$ is a binary relation over $\mathscr{W}$ ("accessibility relation")
- $\mathscr{I}$ is a two-place function that assigns 0 or 1 to each sentence letter, relative to ("at", or "in") each world—that is, for any sentence letter α, and any $w \in \mathscr{W}$, $\mathscr{I}(\alpha, w)$ is either 0 or 1. ("interpretation function")

Each MPL-model contains a set $\mathscr{W}$ of possible worlds, and an accessibility relation $\mathscr{R}$ over $\mathscr{W}$. $\langle \mathscr{W}, \mathscr{R} \rangle$ is sometimes called the model's *frame*. Think of the frame as giving the "structure" of the model's space of possible worlds: it says how many worlds there are, and which worlds are accessible from which. In addition to a frame, each model also contains an interpretation function $\mathscr{I}$, which assigns truth values to sentence letters in worlds.

MPL-models are the configurations for propositional modal logic (recall section 2.2). A configuration is supposed to represent both a way for the world to be, and also the meanings of nonlogical expressions. In MPL-models, the former is represented by the frame. (When we say that a configuration represents "the world", we don't just mean the actual world. "The world" signifies, rather, *reality*, which is here thought of as including the entire space of possible worlds.) The latter is represented by the interpretation function.

(Recall that in propositional logic, the meaning of a sentence letter was a mere truth value. The meaning is now richer: a truth value for each possible world.)

A model's interpretation function assigns truth values only to sentence letters. But the sum total of all the truth values of sentence letters in worlds, together with the frame, determines the truth values of all complex wffs, again relative to worlds. It is the job of the model's valuation function to specify exactly how these truth values get determined:

Definition of valuation: Where $\mathscr{M}$ ($= \langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$) is any MPL-model, the *valuation* for $\mathscr{M}$, $\mathrm{V}_{\mathscr{M}}$, is defined as the two-place function that assigns either 0 or 1 to each wff relative to each member of $\mathscr{W}$, subject to the following constraints, where α is any sentence letter, ϕ and ψ are any wffs, and w is any member of $\mathscr{W}$:

$$\begin{aligned}
\mathrm{V}_{\mathscr{M}}(\alpha, w) &= \mathscr{I}(\alpha, w) \\
\mathrm{V}_{\mathscr{M}}({\sim}\phi, w) &= 1 \text{ iff } \mathrm{V}_{\mathscr{M}}(\phi, w) = 0 \\
\mathrm{V}_{\mathscr{M}}(\phi{\rightarrow}\psi, w) &= 1 \text{ iff either } \mathrm{V}_{\mathscr{M}}(\phi, w) = 0 \text{ or } \mathrm{V}_{\mathscr{M}}(\psi, w) = 1 \\
\mathrm{V}_{\mathscr{M}}(\Box\phi, w) &= 1 \text{ iff for each } v \in \mathscr{W}, \text{if } \mathscr{R}wv, \text{then } \mathrm{V}_{\mathscr{M}}(\phi, v) = 1
\end{aligned}$$

What about truth values for complex formulas containing $\wedge, \vee, \leftrightarrow, \Diamond,$ and $\dashv$? Given the definition of these defined connectives in terms of the primitive connectives, it is easy to prove that the following derived conditions hold:

$$\begin{aligned}
\mathrm{V}_{\mathscr{M}}(\phi\wedge\psi, w) &= 1 \text{ iff } \mathrm{V}_{\mathscr{M}}(\phi, w) = 1 \text{ and } \mathrm{V}_{\mathscr{M}}(\psi, w) = 1 \\
\mathrm{V}_{\mathscr{M}}(\phi\vee\psi, w) &= 1 \text{ iff } \mathrm{V}_{\mathscr{M}}(\phi, w) = 1 \text{ or } \mathrm{V}_{\mathscr{M}}(\psi, w) = 1 \\
\mathrm{V}_{\mathscr{M}}(\phi\leftrightarrow\psi, w) &= 1 \text{ iff } \mathrm{V}_{\mathscr{M}}(\phi, w) = \mathrm{V}_{\mathscr{M}}(\psi, w) \\
\mathrm{V}_{\mathscr{M}}(\Diamond\phi, w) &= 1 \text{ iff for some } v \in \mathscr{W}, \mathscr{R}wv \text{ and } \mathrm{V}_{\mathscr{M}}(\phi, v) = 1 \\
\mathrm{V}_{\mathscr{M}}(\phi\dashv\psi, w) &= 1 \text{ iff for each } v \in \mathscr{W}, \text{ if } \mathscr{R}wv \text{ then either } \mathrm{V}_{\mathscr{M}}(\phi, v) = 0 \text{ or} \\
&\qquad \mathrm{V}_{\mathscr{M}}(\psi, v) = 1
\end{aligned}$$

So far, we have introduced a generic notion of an MPL model, and have defined the notion of a wff's being true at a world in an MPL model. But remember C. I. Lewis's plight: it wasn't clear which modal formulas ought to count as logical truths. His response, and our response, is to construct different modal systems, in which different formulas count as logical truths. The systems we will discuss are named: K, D, T, B, S4, S5. Here in our discussion of semantics, we will come up with different definitions of what counts as a model, one for each system: K, D, T, B, S4, S5. As a result, different formulas will come out valid in the different systems. For example, the formula $\Box P \to \Box\Box P$ is going to come out valid in S4 and S5, but not in the other systems.

The models for the different systems differ according to the formal properties of their accessibility relations. (Formal properties of relations were discussed in section 1.8.) For example, we will define a model for system T ("T-model") as any MPL model whose accessibility relation is reflexive (in $\mathscr{W}$, the set of worlds in that model). Here is the definition:

Definition of model for modal systems: An "S-model", for any of our systems S, is defined as an MPL-model $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$ whose accessibility relation $\mathscr{R}$ has the formal feature given for system S in the following chart:

System	*accessibility relation must be*
K	no requirement
D	serial (in $\mathscr{W}$)
T	reflexive (in $\mathscr{W}$)
B	reflexive (in $\mathscr{W}$) and symmetric
S4	reflexive (in $\mathscr{W}$) and transitive
S5	reflexive (in $\mathscr{W}$), symmetric, and transitive

Thus, *any* MPL-model counts as a K-model, whereas the requirements for the other systems are more stringent.

Our next task is to define validity and semantic consequence for the various systems. A slight wrinkle arises: we can't just define validity as "truth in all models", since formulas aren't simply true or false in MPL-models; they're true or false in various worlds in these models. Instead, we first define a notion of being valid in an MPL model:

Definition of validity in an MPL model: An MPL-wff ϕ is valid in MPL-model $\mathscr{M}$ ($= \langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$ iff for every $w \in \mathscr{W}$, $\mathrm{V}_{\mathscr{M}}(\phi, w) = 1$

Finally we can give the desired definitions:

Definition of validity and semantic consequence:

- An MPL-wff is valid in system S (where S is either K, D, T, B, S4, or S5) iff it is valid in every S-model
- MPL-wff ϕ is a semantic consequence in system S of set of MPL-wffs Γ iff for every S-model $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$ and each $w \in \mathscr{W}$, if $\mathrm{V}_{\mathscr{M}}(\gamma, w) = 1$ for each $\gamma \in \Gamma$, then $\mathrm{V}_{\mathscr{M}}(\phi, w) = 1$

As before, we'll use the $\vDash$ notation for validity and semantic consequence. But since we have many modal systems, if we claim that a formula is valid, we'll need to indicate which system we're talking about. Let's do that by subscripting $\vDash$ with the name of the system; e.g., "$\vDash_{\text{T}} \phi$" means that ϕ is T-valid.

It's important to get clear on the status of possible-worlds lingo here. Where $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$ is an MPL-model, we call the members of $\mathscr{W}$ "worlds", and we call $\mathscr{R}$ the "accessibility" relation. This is certainly a vivid way to talk about these models. But officially, $\mathscr{W}$ is nothing but a nonempty set, any old nonempty set. Its members needn't be the kinds of things metaphysicians call possible worlds. They can be numbers, people, bananas—whatever you like. Similarly for $\mathscr{R}$ and $\mathscr{I}$. The former is just defined to be any old binary relation on $\mathscr{W}$; the latter is just defined to be any old function mapping each pair of a sentence letter and a member of $\mathscr{W}$ to either 1 or 0. Neither needs to have anything to do with the metaphysics of modality. Officially, then, the possible-worlds talk we use to describe our models is just talk, not heavy-duty metaphysics.

Still, models are usually intended to depict some aspect of the real world. The usual intention is that wffs get their truth values within models in a parallel fashion to how natural language sentences are made true by the real world. So if natural language modal sentences aren't made true by anything like possible worlds, then possible worlds semantics would be less valuable than, say, the usual semantics for nonmodal propositional and predicate logic. To be sure, possible worlds semantics would still be useful for various purely formal purposes. For example, given the soundness proofs we will give in section 6.5, the semantics could still be used to establish facts about unprovability in the axiomatic systems to be introduced in section 6.4. But it would be hard to see why possible worlds models would shed any light on the meanings of English modal words, or why truth-in-all-possible-worlds-models would be a good way of modeling (genuine) logical truth for modal statements.

On the other hand, if English modal sentences *are* made true by facts about possible worlds, then the semantics takes on a greater importance. Perhaps then we can, for example, decide what the right logic is, for a given strength of necessity, by reflecting on the formal properties of the accessibility relation—the real accessibility relation, over real possible worlds, not the relation $\mathscr{R}$ over the members of $\mathscr{W}$ in our models. Suppose we're considering some strength, M, of modality. A (real) possible world v is M-accessible from another world, w, iff what happens in v counts as being M-possible, from the point of view of w. Perhaps we can figure out the logic of M-necessity and M-possibility by investigating the formal properties of M-accessibility.

Consider deontic necessity and possibility, for example: a proposition is deontically necessary iff it ought to be the case; a proposition is deontically possible iff it is morally acceptable that it be the case. The relation of deontic accessibility seems not to be reflexive: in an imperfect world like our own, many things that ought not to be true are nevertheless true. Thus, a world can fail to be deontically accessible relative to itself. (As we will see, this corresponds to the fact that deontic necessity is non-alethic; it does not imply truth.) On the other hand, one might argue, deontic accessibility *is* serial, since surely there must always be *some* deontically accessible world—some world in which what occurs is morally acceptable. (To deny this would be to admit that everything could be forbidden.) So, perhaps system D gives the logic of deontic necessity and possibility (see also section 7.1).

To take one other example: some have argued that the relation of *metaphysical*-accessibility (the relation relevant to metaphysical necessity and possibility) is a total relation: every world is metaphysically possible relative to every other.[5] What modal logic would result from requiring $\mathscr{R}$ to be a total (in $\mathscr{W}$) relation? The answer is: S5. That is, you get the same valid formulas whether you require $\mathscr{R}$ to be a total relation or an equivalence relation (see exercise 6.1). So, if the (real) metaphysical accessibility relation is a total relation, the correct logic for metaphysical necessity is S5. But others have argued that metaphysical accessibility is intransitive.[6] Perhaps one possible world is metaphysically accessible from another only if the individuals in the latter world aren't too different from how they are in the former world—only if such differences are below a certain threshold. In that case, it might be argued, a world in which I'm a frog is not metaphysically accessible from the actual world: any world in which I'm that drastically different from my actual, human, self, just isn't metaphysically possible, relative to actuality. But perhaps a world, w, in which I'm a human-frog hybrid *is* accessible from the actual world (the difference between a human and a frog-human hybrid is below the threshold); and perhaps the frog world is accessible from w (since the difference between a frog-human hybrid and a frog is also below the threshold). If so, then metaphysical accessibility is intransitive. Metaphysical accessibility is clearly reflexive. So perhaps the logic of metaphysical possibility is given by system B or system T.

[5] See Lewis (1986, 246).

[6] Compare Salmon (1986).

Exercise 6.1** Let O be the modal system given by the requirement that $\mathscr{R}$ must be total (in $\mathscr{W}$). Show that $\vDash_{\mathrm{O}} \phi$ iff $\vDash_{\mathrm{S5}} \phi$.

6.3.2 Semantic validity proofs

Given our definitions, we can now show particular formulas to be valid in various systems.

Example 6.1: The wff $\Box(P\vee{\sim}P)$ is K-valid. To show this, we must show that the wff is valid in all MPL-models, since validity-in-all-MPL-models is the definition of K-validity. Being valid in a model means being true at every world in the model. So, consider any MPL-model $\langle\mathscr{W},\mathscr{R},\mathscr{I}\rangle$, and let w be any world in $\mathscr{W}$. We must show that $\mathrm{V}_{\mathscr{M}}(\Box(P\vee{\sim}P),w)=1$. (As before, I'll start to omit the subscript $\mathscr{M}$ on $\mathrm{V}_{\mathscr{M}}$ when it's clear which model we're talking about.)

i) Suppose for reductio that $\mathrm{V}(\Box(P\vee{\sim}P),w)=0$

ii) So, by the truth condition for $\Box$ in the definition of the valuation function, there is some world, v, such that $\mathscr{R}wv$ and $\mathrm{V}(P\vee{\sim}P,v)=0$

iii) Given the (derived) truth condition for $\vee$, $\mathrm{V}(P,v)=0$ and $\mathrm{V}({\sim}P,v)=0$

iv) Since $\mathrm{V}({\sim}P,v)=0$, given the truth condition for $\sim$, $\mathrm{V}(P,v)=1$. But that's impossible; $\mathrm{V}(P,v)$ can't be both 0 and 1.

Thus, $\vDash_{\mathrm{K}} \Box(P\vee{\sim}P)$.

Note that similar reasoning would establish $\vDash_{\mathrm{K}} \Box\phi$, for any tautology ϕ. For within any world, the truth values of complex statements of propositional logic are determined by the truth values of their constituents in that world by the usual truth tables. So if ϕ is a tautology, it will be true in any world in any model; hence $\Box\phi$ will turn out true in any world in any model.

Example 6.2: Show that $\vDash_{\mathrm{T}} (\Diamond\Box(P\rightarrow Q)\wedge\Box P)\rightarrow\Diamond Q$. Let w be any world in any T-model $\mathscr{M}$; we must show that $\mathrm{V}_{\mathscr{M}}((\Diamond\Box(P\rightarrow Q)\wedge\Box P)\rightarrow\Diamond Q,w)=1$:

i) Suppose for reductio that $\mathrm{V}((\Diamond\Box(P\rightarrow Q)\wedge\Box P)\rightarrow\Diamond Q,w)=0$.

ii) So $\mathrm{V}(\Diamond\Box(P\rightarrow Q)\wedge\Box P,w)=1$ and …

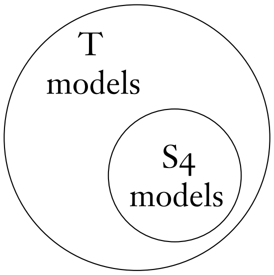
iii) $\dots \mathrm{V}(\Diamond Q, w) = 0$. So Q is false in every world accessible from w.

iv) From ii), $\Diamond\Box(P\rightarrow Q)$ is true at w, and so $\mathrm{V}(\Box(P\rightarrow Q), v) = 1$, for some world, call it v, such that $\mathscr{R}wv$.

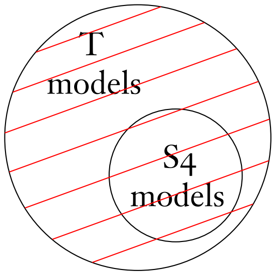
v) From ii), $\mathrm{V}(\Box P, w) = 1$. So, by the truth condition for the $\Box$, P is true in every world accessible from w; since $\mathscr{R}wv$, it follows that $\mathrm{V}(P, v) = 1$. But $\mathrm{V}(Q, v) = 0$ given iii). So $\mathrm{V}(P\rightarrow Q) = 0$.

vi) From iv), $P\rightarrow Q$ is true in every world accessible from v; since $\mathscr{M}$ is a T-model, $\mathscr{R}$ is reflexive; so $\mathscr{R}vv$; so $\mathrm{V}(P\rightarrow Q, v) = 1$, contradicting v).

The last example showed that the formula $(\Diamond\Box(P\rightarrow Q)\wedge\Box P) \rightarrow \Diamond Q$ is valid in T. Suppose we wanted to show that it is also valid in S4. What more would we have to do? Nothing! To be S4-valid is to be valid in every S4-model. But a quick look at the definitions shows that every S4-model is a T-model. So, since we already know that the the formula is valid in all T-models, we may conclude that it must be valid in all S4-models without doing a separate proof:



The S4-models are a subset of the T-models.



So if a formula is valid in all T-models, it's automatically valid in all S4-models

Think of it another way. A proof that a wff is S4-valid *may* use the information that the accessibility relation is both transitive and reflexive. But it doesn't need to. So the T-validity proof in example 6.2 also counts as an S4-validity proof. (It also counts as a B-validity proof and an S5-validity proof.) But it doesn't count as a K-validity proof, since it assumes in line vi) that $\mathscr{R}$ is reflexive. To be K-valid, a wff must be valid in *all* models, whereas the proof in example 6.2 only establishes validity in all reflexive models. (In fact $(\Diamond\Box(P\rightarrow Q)\wedge\Box P) \rightarrow \Diamond Q$ isn't K-valid, as we'll be able to demonstrate shortly.)

Consider the following diagram of systems:

$$\begin{tikzcd}
 & S5 & \\
S4 \arrow[ur] & & B \arrow[ul] \\
 & T \arrow[ul] \arrow[ur] & \\
 & D \arrow[u] & \\
 & K \arrow[u] &
\end{tikzcd}$$

An arrow from one system to another indicates that validity in the first system implies validity in the second system. For example, all D-valid wffs are also T-valid. For if a wff is valid in all D-models, then, since every T-model is also a D-model (reflexivity implies seriality), it must be valid in all T-models as well.

S5 is the strongest system, since it has the most valid formulas. That's because it has the fewest models: it's easy to be S5-valid since there are so few potentially falsifying models. K is the weakest system—fewest validities—since it has the most potentially falsifying models. The other systems are intermediate.

Notice that the diagram isn't linear. Both B and S4 are stronger than T: each contains all the T-valid formulas and more besides. And S5 is stronger than both B and S4. But (as we will see below) neither B nor S4 is stronger than the other (nor are they equally strong): some B-valid wffs aren't S4-valid, and some S4-valid wffs aren't B-valid. (The definitions of B and S4 hint at this. B requires symmetry but not transitivity, whereas S4 requires transitivity but not symmetry, so some B-models aren't S4-models, and some S4-models aren't B-models.)

Suppose you're given a formula, and for each system in which it is valid, you want to give a semantic proof of its validity. This needn't require multiple semantic proofs. As we saw with example 6.2, to prove that a wff is valid in a number of systems, it suffices to give a validity proof in the weakest of those systems, since that very proof will automatically be a proof that it is valid in all stronger systems. For example, a K-validity proof is itself a validity proof for D, T, B, S4, and S5. But there is an exception. Suppose a wff is *not* valid in T, but you've given a semantic proof of its validity in B. This proof also shows that the wff is S5-valid, since every S5-model is a B-model. But you can't yet conclude that the wff is S4-valid, since not every S4-model is a B-model. Another semantic proof may be needed: of the formula's S4-validity. (Of course, the formula may not be S4-valid.) So: when a wff is valid in both B and S4, but not in T, two semantic proofs of its validity are needed.

We are now in a position to do validity proofs. But as we'll see in the next section, it's often easier to do proofs of validity when one has failed to construct a counter-model for a formula.

Exercise 6.2 Use validity proofs to demonstrate the following:

a) $\vDash_{\mathrm{D}} [\Box P \wedge \Box({\sim} P \vee Q)] \rightarrow \Diamond Q$

b) $\vDash_{\mathrm{S4}} \Diamond\Diamond(P \wedge Q) \rightarrow \Diamond Q$

6.3.3 Countermodels

We have a definition of validity for the various systems, and we've shown how to establish validity of particular formulas. (We have also defined semantic consequence for these systems, but our focus will be on validity.) Now we'll see how to establish *in*validity. We establish that a formula is invalid by constructing a countermodel for it—a model containing a world in which the formula is false. (Since validity means truth in every world in every model, the existence of a single countermodel establishes invalidity.)

I'm going to describe a helpful graphical procedure, introduced by Hughes and Cresswell (1996), for constructing countermodels. Now, it's always an option to bypass the graphical procedure and directly intuit what a counter-model might look like. But the graphical procedure makes things a lot easier, especially with more complicated formulas.

I'll illustrate the procedure by using it to show that the wff $\Diamond P \rightarrow \Box P$ is *not* K-valid. To be K-valid, a wff must be valid in all MPL-models, so all we must do is find one MPL-model in which $\Diamond P \rightarrow \Box P$ is false in some world.

Place the formula in a box

We begin by drawing a box, which represents some chosen world in the model we're in the process of pictorially constructing. The goal is to make the formula false in this world. In these examples I'll always call this first world "r":

$$\mathrm{r}\ \boxed{\Diamond P\to\Box P}$$

Now, since the box represents a world, we should have some way of representing the accessibility relation. What worlds are accessible from r; what worlds does r "see"? Well, to represent one world (box) seeing another, we'll draw an arrow from the first to the second. But in this case we don't need to draw any arrows. We're only trying to show that $\Diamond P\to\Box P$ is K-invalid, and the accessibility relation for system K doesn't even need to be serial—no world needs to see any worlds at all. So, we'll forget about arrows for the time being.

Make the formula false in the world

We'll indicate a formula's truth value by writing that truth value above the formula's major connective. (The "major connective" of a wff is the last connective that was added when the wff was formed via the rules of grammar.[7] Thus, the major connective of $P\to\Box Q$ is the →, and the major connective of $\Box(P\to\Box Q)$ is the leftmost □.) So to indicate that $\Diamond P\to\Box P$ is to be false in this model, we'll put a 0 above its arrow:

$$\mathrm{r}\ \boxed{\Diamond P\overset{0}{\to}\Box P}$$

Enter forced truth values

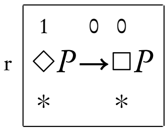
Assigning a truth value to a formula sometimes forces us to assign truth values to other formulas in the same world. For example, if we make a conjunction true in a world then we must make each of its conjuncts true at that world; and if we make a conditional false at a world, we must make its antecedent true and its consequent false at that world. In the current example, since we've made $\Diamond P\to\Box P$ false in r, we've got to make $\Diamond P$ true at r (indicated on the diagram by a 1 over its major connective, the ◇), and we've got to make its consequent $\Box P$ false at r:

$$\mathrm{r}\ \boxed{\overset{1}{\Diamond} P\overset{0}{\to}\overset{0}{\Box} P}$$

[7] In talking about major connectives, let's treat nonprimitive connectives as if they were primitive. Thus, the major connective of $\Box P\wedge{\sim}Q$ is the ∧.

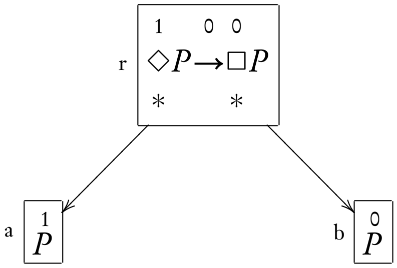
Enter asterisks

When we assign a truth value to a modal formula, we thereby commit ourselves to assigning certain other truth values to various formulas at various worlds. For example, when we make $\Diamond P$ true at r, we commit ourselves to making P true at some world that r sees. To remind ourselves of this commitment, we'll put an asterisk (*) below $\Diamond P$. An asterisk *below* indicates a commitment to there being *some* world of a certain sort. Similarly, since $\Box P$ is false at r, this means that P must be false in some world P sees (if it were true in all such worlds then $\Box P$ would be true at r). We again have a commitment to there being some world of a certain sort, so we enter an asterisk below $\Box P$ as well:



Discharge bottom asterisks

The next step is to fulfill the commitments we incurred when we added the bottom asterisks. For each, we need to add a world to the diagram. The first asterisk requires us to add a world in which P is true; the second requires us to add a world in which P is false. We do this as follows:



The official model

We now have a diagram of a K-model containing a world in which $\Diamond P \rightarrow \Box P$ is false. But we need to produce an official model, according to the official definition of a model. A model is an ordered triple $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$, so we must specify the model's three members.

The set of worlds, $\mathscr{W}$, is simply the set of worlds I invoked:

$$\mathscr{W} = \{\mathrm{r}, \mathrm{a}, \mathrm{b}\}$$

What *are* r, a, and b? Let's just take them to be the letters 'r', 'a', and 'b'. No reason not to—the members of $\mathscr{W}$, recall, can be any things whatsoever.

Next, the accessibility relation. This is represented on the diagram by the arrows. In our model, there is an arrow from r to a, an arrow from r to b, and no other arrows. Thus, the diagram represents that r sees a, that r sees b, and that there are no further cases of seeing. Now, remember that the accessibility relation, like all relations, is a set of ordered pairs. So, we simply write out this set:

$$\mathscr{R} = \{\langle \mathrm{r}, \mathrm{a}\rangle, \langle \mathrm{r}, \mathrm{b}\rangle\}$$

That is, we write out the set of all ordered pairs $\langle w_1, w_2\rangle$ such that w_1 "sees" w_2.

Finally, we need to specify the interpretation function, $\mathscr{I}$, which assigns truth values to sentence letters at worlds. In our model, $\mathscr{I}$ must assign 1 to P at world a, and 0 to P at world b. Now, our official definition requires an interpretation to assign a truth value to each of the infinitely many sentence letters at each world; but so long as P is true at world a and false at world b, it doesn't matter what other truth values $\mathscr{I}$ assigns. So let's just (arbitrarily) choose to make all other sentence letters false at all worlds in the model. We have, then:

$$\mathscr{I}(P, \mathrm{a}) = 1$$
$$\mathscr{I}(P, \mathrm{b}) = 0$$
$$\mathscr{I}(\alpha, w) = 0 \text{ for all other sentence letters } \alpha \text{ and worlds } w$$

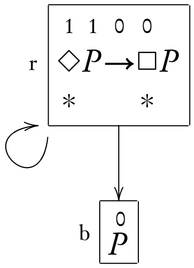
That's it—we're done. We have produced a model in which $\Diamond P \rightarrow \Box P$ is false at some world; hence this formula is not valid in all models; and hence it's not K-valid: $\nvDash_{\mathrm{K}} \Diamond P \rightarrow \Box P$.

Check the model

At the end of this process, it's a good idea to double-check that your model is correct. This involves various things. First, make sure that you've succeeded in producing the correct kind of model. For example, if you're trying to produce a T-model, make sure that the accessibility relation you've written down is reflexive. (In our case, we were only trying to construct a K-model, and so for us this step is trivial.) Second, make sure that the formula in question really does come out false at one of the worlds in your model.

Simplifying models

Sometimes a model can be simplified. In the countermodel for $\Diamond P \rightarrow \Box P$, we needn't have used three worlds. We added world a because the truth of $\Diamond P$ called for a world that r sees in which P is true. But we needn't have made that a *new* world—we could have made P true in r and made r see itself. (We couldn't have done that for both asterisks; that would have made P both true and false at r.) So, we could make this one simplification:

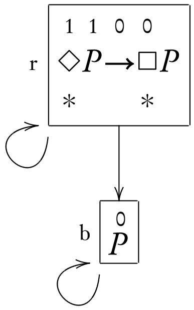


Official model:

$$\mathscr{W} = \{r, b\}$$
$$\mathscr{R} = \{\langle r, r\rangle, \langle r, b\rangle\}$$
$$\mathscr{I}(P, \mathrm{r}) = 1, \text{ all others } 0$$

Adapting models to different systems

We have shown that $\Diamond P \rightarrow \Box P$ is not K-valid. Next let's show that this formula isn't D-valid—that it is false in some world of some model with a serial accessibility relation. The model we just constructed won't do, since its accessibility relation isn't serial; world b doesn't see any world. But we can easily change that:

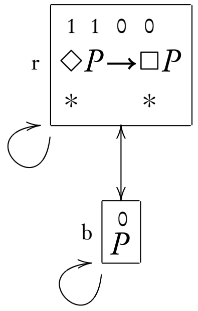


Official model:

$$\mathscr{W} = \{r, b\}$$
$$\mathscr{R} = \{\langle r, r\rangle, \langle r, b\rangle, \langle b, b\rangle\}$$
$$\mathscr{I}(P, \mathrm{r}) = 1, \text{ all others } 0$$

That was easy—adding the fact that b sees itself didn't require changing anything else in the model.

Suppose we want now to show that $\Diamond P \rightarrow \Box P$ isn't T-valid. What more must we do? Nothing! The model we just displayed is a T-model, in addition to being a D-model, since its accessibility relation is reflexive. In fact, its accessibility relation is also transitive, so it's also an S4-model. What about B? It's easy to make the accessibility relation symmetric:



Official model:

$$\mathscr{W} = \{r, b\}$$
$$\mathscr{R} = \{\langle \mathrm{r},\mathrm{r}\rangle, \langle \mathrm{r},\mathrm{b}\rangle, \langle \mathrm{b},\mathrm{b}\rangle, \langle \mathrm{b},\mathrm{r}\rangle\}$$
$$\mathscr{I}(P,\mathrm{r}) = 1, \text{ all others } 0$$

So we've established B-invalidity as well. In fact, the model just displayed is an S5-model since its accessibility relation is an equivalence relation. And so, since any S5-model is also a K, D, T, B, and S4-model, this one model shows that $\Diamond P \to \Box P$ is not valid in *any* of our systems. So we have established that: $\nvDash_{\mathrm{K,D,T,B,S4,S5}} \Diamond P \to \Box P$.

In this case it wouldn't have been hard to move straight to the final S5-model, right from the start. But in more difficult cases, it's best to proceed slowly, as I did here. Try first for a countermodel in K. Then build the model up gradually, trying to make its accessibility relation satisfy the requirements of stronger systems. When you get a countermodel in a stronger system (a system with more requirements on its models), that very countermodel will establish invalidity in all weaker systems. Keep in mind the diagram of systems:

$$\begin{tikzcd}
 & S5 & \\
S4 \arrow[ur] & & B \arrow[ul] \\
 & T \arrow[ul] \arrow[ur] & \\
 & D \arrow[u] & \\
 & K \arrow[u] &
\end{tikzcd}$$

An arrow from one system to another, recall, indicates that validity in the first system implies validity in the second. The arrows also indicate facts about *invalidity*, but in reverse: when an arrow points from one system to another, then invalidity in the *second* system implies invalidity in the *first*. For example, if a wff is invalid in T, then it is invalid in D. (That's because every T-model is a D-model; a countermodel in T is therefore a countermodel in D.)

When our task is to discover the systems in which a given formula is invalid, usually only one countermodel will be needed—a countermodel in the strongest system in which the formula is invalid. But there is an exception involving B and S4. Suppose a given formula is valid in S5, but we discover a model showing that it isn't valid in B. That model is automatically a T, D, and K-model, so we know that the formula isn't T, D, or K-valid. But we don't yet know about S4-validity. If the formula is S4-invalid, then we will need to produce a second countermodel, an S4 countermodel. (Notice that the B-model couldn't *already* be an S4-model. If it were, then its accessibility relation would be reflexive, symmetric, and transitive, and so it would be an S5-model, contradicting the fact that the formula was S5-valid.)

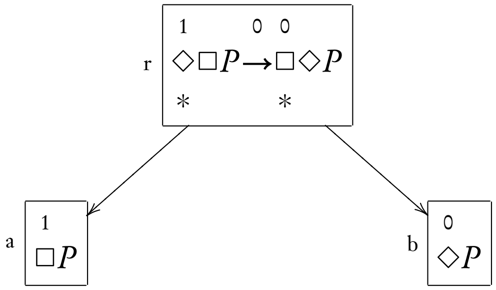
So far we have the following steps for constructing countermodels:

1. Place the formula in a box and make it false
2. Enter forced truth values
3. Enter asterisks
4. Discharge bottom asterisks
5. The official model

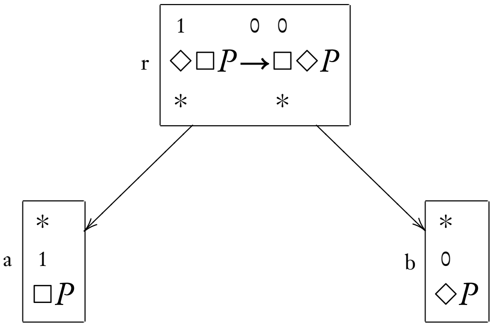
We need to add to this list.

Top asterisks

Let's try to get a countermodel for $\Diamond\Box P\rightarrow\Box\Diamond P$ in all the systems in which it is invalid. A cautious beginning would be to try for a K-model. After the first few steps, we have:

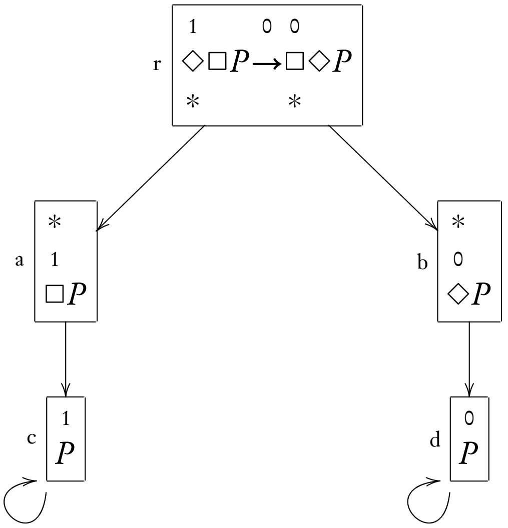


At this point we have a true $\Box$ (in world a) and a false $\Diamond$ (in world b). Like true $\Diamond$s and false $\Box$s, these generate commitments pertaining to other worlds. But unlike true $\Diamond$s and false $\Box$s, they don't commit us to the existence of *some* accessible world of a certain type; they carry commitments for *every* accessible world. The true $\Box P$ in world a, for example, requires us to make P true in every world accessible from a. Similarly, the falsity of $\Diamond P$ in world b commits us to making P false in every world accessible from b. We indicate such commitments, universal rather than existential, by putting asterisks *above* the relevant modal operators:



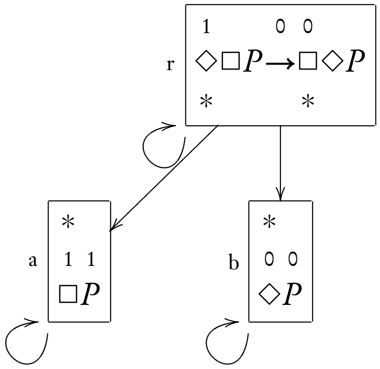
Now, how can we honor these commitments; how must we "discharge" these asterisks? In this case, when trying to construct a K-model, we don't need to do anything. Since world a, for example, doesn't see any world, P is automatically true in every world it sees; the statement "for every world, w, if $\mathscr{R}aw$ then $\mathrm{V}(P,w)=1$" is vacuously true. Same goes for b—P is automatically false in all worlds it sees. So, we've got a K-model in which $\Diamond\Box P\rightarrow\Box\Diamond P$ is false.

Now let's turn the model into a D-model. Every world must now see at least one world. Let's try:



I added worlds c and d, so that a and b would each see at least one world. (Further, worlds c and d each had to see a world, to keep the relation serial. I could have added new worlds e and f seen by c and d, but e and f would have needed to see some worlds. So I just let c and d see themselves.) But once c and d were added, discharging the upper asterisks in worlds a and b required making *P* true in c and false in d (since a sees c and b sees d).

Let's now try for a T-model. Worlds a and b must now see themselves. But then we no longer need worlds c and d, since they were added just to make the relation serial. So we can simplify:



Official model:

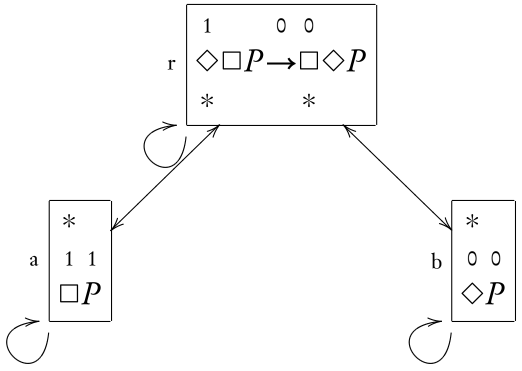
$$\mathscr{W} = \{\mathrm{r},\mathrm{a},\mathrm{b}\}$$
$$\mathscr{R} = \{\langle \mathrm{r},\mathrm{r}\rangle,\langle \mathrm{a},\mathrm{a}\rangle,\langle \mathrm{b},\mathrm{b}\rangle,\langle \mathrm{r},\mathrm{a}\rangle,\langle \mathrm{r},\mathrm{b}\rangle\}$$
$\mathscr{I}(P,\mathrm{a}) = 1$, all others 0

When you add arrows, you need to make sure that all top asterisks are discharged. In this case this required nothing of world r, since there were no top asterisks there. There were top asterisks in worlds a and b; these I discharged by making P be true in a and false in b.

Notice that I could have moved straight to this T-model—which is itself a D-model—rather than first going through the earlier mere D-model. However, this won't always be possible—sometimes you'll be able to get a D-model, but no T-model.

At this point let's verify that our model does indeed assign the value 0 to our formula $\Diamond\Box P\rightarrow\Box\Diamond P$. First notice that $\Box P$ is true in a (since a only sees one world—itself—and P is true there). But r sees a. So $\Diamond\Box P$ is true at r. Now, consider b. b sees only one world, itself; and P is false there. So $\Diamond P$ must also be false there. But r sees b. So $\Box\Diamond P$ is false at r. But now, the antecedent of $\Diamond\Box P\rightarrow\Box\Diamond P$ is true, while its consequent is false, at r. So that conditional is false at r. Which is what we wanted.

Onward. Our model is not a B-model since r sees a and b but they don't see r back. Suppose we try to make a and b see r:



We must now make sure that all top asterisks are discharged. Since a now sees r, P must be true at r. But b sees r too, so P must be false at r. Since P can't be both true and false at r, we're stuck. We have failed to construct a B-model in which this formula is false.

Our failure to construct a B-countermodel suggests that it may be impossible to do so. We can *prove* that this is impossible by showing that the formula is true in every world of every B-model—that is, that the formula is B-valid. Let $\mathcal{M}=\langle\mathcal{W},\mathcal{R},\mathcal{I}\rangle$ be any model in which $\mathcal{R}$ is reflexive and symmetric, and consider any $w\in\mathcal{W}$; we must show that $V_{\mathcal{M}}(\Diamond\Box P\rightarrow\Box\Diamond P,w)=1$:

i) Suppose for reductio that $V(\Diamond\Box P\rightarrow\Box\Diamond P,w)=0$. Then $V(\Diamond\Box P,w)=1$ and $V(\Box\Diamond P,w)=0$.

ii) Given the former, for some v, $\mathscr{R}wv$ and $\mathrm{V}(\Box P, v) = 1$.

iii) Given the latter, for some u, $\mathscr{R}wu$ and $\mathrm{V}(\Diamond P, u) = 0$.

iv) From ii), P is true at every world accessible from v; by symmetry, $\mathscr{R}vw$; so $\mathrm{V}(P, w) = 1$.

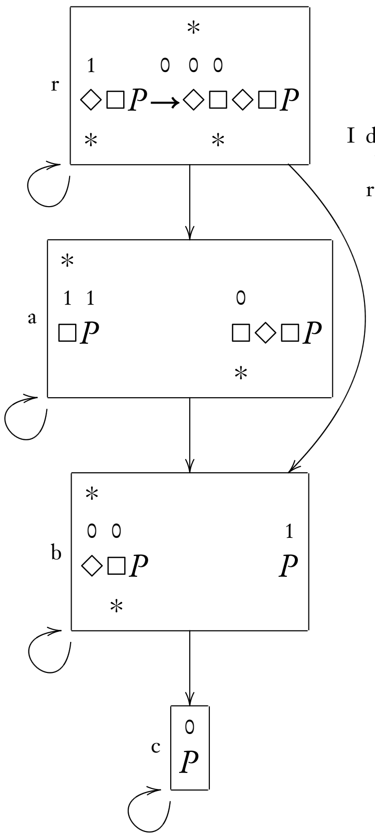
v) From iii), P is false at every world accessible from u; by symmetry, $\mathscr{R}uw$; so $\mathrm{V}(P, w) = 0$, contradicting iv)

Just as we suspected: the formula is indeed B-valid; no wonder we failed to come up with a B-countermodel!

Might there be an S5 countermodel? No: the B-validity proof we just constructed also shows that the formula is S5-valid. What about an S4 countermodel? The existence of the B-validity proof doesn't tell us one way or the other. Remember the diagram: validity in S4 doesn't imply validity in B, nor does validity in B imply validity in S4. So we must either try to come up with an S4-model, or try to construct an S4 semantic validity proof. Usually it's best to try for a model. In the present case this is easy: the T-model we gave earlier is itself an S4-model. Thus, on the basis of that model, we can conclude that $\nvDash_{\mathrm{K,D,T,S4}} \Diamond\Box P \rightarrow \Box\Diamond P$.

We have accomplished our task. We gave an S4 countermodel, which is a countermodel for each system in which $\Diamond\Box P \rightarrow \Box\Diamond P$ is invalid. And we gave a validity proof in B, which is a validity proof for each system in which the formula is valid.

Example 6.3: Determine in which systems $\Diamond\Box P \rightarrow \Diamond\Box\Diamond\Box P$ is valid and in which systems it is invalid. We can get a T-model as follows:



Notice how commitments to truth values for different formulas are recorded by placing the formulas side by side in the box

I discharged the second bottom asterisk in r by letting r see b

Official model:

$$\mathscr{W} = \{r, a, b, c\}$$
$$\mathscr{R} = \{\langle r,r\rangle, \langle a,a\rangle, \langle b,b\rangle, \langle c,c\rangle, \langle r,a\rangle, \langle r,b\rangle, \langle a,b\rangle, \langle b,c\rangle\}$$
$$\mathscr{I}(P,\mathrm{a}) = \mathscr{I}(P,\mathrm{b}) = 1, \text{ all others } 0$$

Now consider what happens when we try to turn this model into a B-model. World b must see back to world a. But then the false $\Diamond\Box P$ in b conflicts with the true $\Box P$ in a. So it's time for a validity proof. In constructing this validity proof, we can be guided by our failed attempt to construct a countermodel (assuming all of our choices in constructing that countermodel were forced). In the following proof that the formula is B-valid, I use variables for worlds that match up with the attempted countermodel above:

i) Suppose for reductio that $\mathrm{V}(\Diamond\Box P\rightarrow\Diamond\Box\Diamond\Box P, r) = 0$, in some world r in some B-model $\langle \mathscr{W}, \mathscr{R}, \mathscr{I}\rangle$. So $\mathrm{V}(\Diamond\Box P, r) = 1$ and $\mathrm{V}(\Diamond\Box\Diamond\Box P, r) = 0$.

ii) Given the former, for some world a, $\mathscr{R}ra$ and $\text{V}(\Box P, a) = 1$.

iii) Given the latter, since $\mathscr{R}ra$, $\text{V}(\Box\Diamond\Box P, a) = 0$. So for some b, $\mathscr{R}ab$ and $\text{V}(\Diamond\Box P, b) = 0$. By symmetry, $\mathscr{R}ba$; so $\text{V}(\Box P, a) = 0$, contradicting ii).

We now have a T-model for the formula, and a proof that it is B-valid. The B-validity proof shows the formula to be S5-valid; the T-model shows it to be K- and D-invalid. We don't yet know about S4. So let's return to the T-model above and try to make its accessibility relation transitive. World a must then see world c, which is impossible since $\Box P$ is true in a and P is false in c. So we're ready for a S4-validity proof (the proof looks like the B-validity proof at first, but then diverges):

i) Suppose for reductio that $\text{V}(\Diamond\Box P \to \Diamond\Box\Diamond\Box P, r) = 0$, in some world r in some B-model $\langle \mathscr{W}, \mathscr{R}, \mathscr{I} \rangle$. So $\text{V}(\Diamond\Box P, r) = 1$ and $\text{V}(\Diamond\Box\Diamond\Box P, r) = 0$.

ii) Given the former, for some world a, $\mathscr{R}ra$ and $\text{V}(\Box P, a) = 1$.

iii) Given the latter, since $\mathscr{R}ra$, $\text{V}(\Box\Diamond\Box P, a) = 0$. So for some b, $\mathscr{R}ab$ and $\text{V}(\Diamond\Box P, b) = 0$. By reflexivity, $\mathscr{R}bb$, so $\text{V}(\Box P, b) = 0$. So for some world c, $\mathscr{R}bc$ and $\text{V}(P, c) = 0$.

iv) Since $\mathscr{R}ab$ and $\mathscr{R}bc$, by transitivity we have $\mathscr{R}ac$. So, given ii), $\text{V}(P, c) = 1$, contradicting iii)