9
First-order logic revisited

We now take time off from computability to review some logic, after which we'll apply what we've learned about computability to logic and show, among other things, that there can be no perfectly satisfactory uniform mechanical procedure for determining whether or not inferences in the notation of first-order logic (= elementary logic) are valid.

We assume that you are already familiar with logical notation: that you've seen the *connectives*

–	&	v	→	↔
not	*and*	*or*	*if ... then*	*if and only if*

and the quantifiers

$\forall x \quad \forall y \quad \forall z \quad \ldots$ *universal quantifiers*

$\exists x \quad \exists y \quad \exists z \quad \ldots$ *existential quantifiers*

either in the notation illustrated above or in some other. We assume that once you are informed that the universe of discourse is mankind, that '*L*' means 'loves', and that '*a*' names Alma, you can read such formulas as

$$\forall x(xLx \rightarrow \exists y\, xLy). \tag{9.1}$$

('Anyone who loves himself loves someone') and

$$\forall x[\exists y(xLy \,\&\, yLa) \rightarrow -xLa]. \tag{9.2}$$

('None of Alma's lovers' lovers love her'). We also assume that with a bit of thought you can recognize (9.1) as a logical truth (= as valid) and that you can recognize that (9.2) implies '$-aLa$' (= the inference from (9.2) as premise to '$-aLa$' as conclusion is a *valid* one).

You are probably familiar, too, with the use of the sign

$$=$$

of identity (= the equals-sign) in making such statements as

$$\forall x\, \forall y\, \forall z\, [(xFz \,\&\, yFz) \rightarrow x = y] \tag{9.3}$$

which (reading '*F*' as *is a father of*) says that nobody has more than one father. You may be familiar, too, with the use of function symbols such as '*f*' for *the father of* in making such statements as

$$aLf(f(a)) \tag{9.4}$$

('Alma loves her paternal grandfather') which can also be made, more awkwardly, without function symbols.

In this chapter we will provide a framework within which such notions as validity can be discussed with the degree of clarity that will be needed when we show that the valid inferences are precisely those that pass a certain mechanical test. (We'll also show that there can be no such mechanical test for invalidity.)

The first notion we shall need is a division of the symbols that may occur in formulas in the notation of elementary logic into two sorts: *logical* and *non-logical*. The logical symbols are the variables and these eleven: – & v → ↔ ∃ ∀ = ( ) ,

We stress that the equals-sign counts as a *logical* symbol: we shall give it special treatment when we come to state the conditions under which sentences in logical notation are true or false. We suppose that there are enumerably infinitely many variables.

The non-logical symbols are of four disjoint sorts: *names*, *function symbols*, *sentence letters*, and *predicate letters*. Both function symbols and predicate letters are of various (unique) *numbers of places*: thus, ordinarily, + is a *two*-place function symbol and < is a *two*-place predicate letter. Any positive integer can be the number of places of a predicate letter or a function symbol. (Occasionally names are regarded as zero-place function symbols and sentence letters as zero-place predicate letters. But though this is sometimes convenient, we shall not usually regard them in this way.)

The equals-sign, even though it is a logical symbol, is to count as a two-place predicate letter. But it is the only exception: the non-logical symbols are precisely the names, function symbols, sentence letters, and predicate letters other than =.

We shall appropriate the word 'language' to mean *an enumerable set of non-logical symbols*. A sentence or a formula *of* or *in* a language is one whose non-logical symbols all belong to the language. We suppose that sentences and formulas are formed in the ordinary, familiar ways, and that a sentence is, as usual, a formula in which there are no *free* occurrences of any variables. As always, an occurrence of a variable in a formula is free if it is governed by no quantifier (in that formula) containing that variable.

One language will be of great interest to us in later chapters. It is called *the language of arithmetic* – its nickname is 'L' – and is the set $\{\mathbf{o}, {}', +, \cdot\}$. $\mathbf{o}$ is a name, $'$ is a one-place function symbol, and $+$ and $\cdot$ are two-place function symbols. There are no sentence letters or predicate letters in L. Two of the enumerably infinitely many sentences of L are

$$\mathbf{o}''+\mathbf{o}''' = \mathbf{o}''''' \tag{9.5}$$

and
$$\forall x \forall y\, x+y = y+x. \tag{9.6}$$

((9.5) and (9.6) are really a sort of slang for what the official conventions about the positioning of function symbols and predicate letters would have us write, namely,

$$=(+({}'({}'(\mathbf{o})), {}'({}'({}'(\mathbf{o}))), {}'({}'({}'({}'({}'(\mathbf{o})))))))$$

and
$$\forall x \forall y = (+(x, y), +(y, x)).$$

We shall continue to speak with the vulgar.)

The empty set, $\varnothing$, is certainly an enumerable set of non-logical symbols, and so, by our lights, it is a language. One might be tempted to suppose that there weren't any sentences of this language, but there are: as $=$ is a predicate letter

$$\forall x\, x = x$$

is one of its sentences, and indeed is a sentence of *every* language.

Any formula in a language is a finite sequence drawn from an enumerable set of symbols. So there are at most enumerably many formulas in any language. (Cf. Exercise 2.4(d).) But as

$$\forall x\, x = x, \quad \forall y\, y = y, \quad \forall z\, z = z, \ldots$$

are sentences in every language, the set of sentences and hence the set of formulas of any language are actually enumerably infinite.

As yet we haven't said anything about when sentences in various languages are either true or false. *Interpretations* (or *models* or *structures*, as they are sometimes called) remedy this lack. An *interpretation of a language* specifies these things:

(1) A *domain*, *viz.* a nonempty set. The domain of an interpretation is the range (according to the interpretation) of any variables that occur in any sentences in the language. ('Universe' and 'universe of discourse' are often used as synonyms of 'domain'.)

(2) For each name in the language, and no others, a *designation* (or *bearer*, or *denotation*, or *reference*), i.e. an object in the domain specified in (1).

(3) For each function symbol in the language and no others, a *function* $\mathfrak{f}$ which assigns a value in the domain for any sequence of arguments in the domain; in the case of an n-place function symbol, f, the function $\mathfrak{f}$ has n argument places. Thus if f is a one-place function symbol, the domain of $\mathfrak{f}$ (the set of arguments of $\mathfrak{f}$) will be just the domain of the interpretation, and the range of $\mathfrak{f}$ (the set of values of $\mathfrak{f}$) will be a subset of the domain of the interpretation.

(4) For each sentence letter in the language and no others, a *truth-value*, 1 (truth) or 0 (falsity). (Truth is one.)

(5) For each predicate letter in the language and no others, a *characteristic function*. In the case of an n-place predicate letter R, the characteristic function ϕ has n argument places. For any objects $o_1, \ldots, o_n$ in the domain specified in (1), the value $\phi(o_1, \ldots, o_n)$ is to be 1 or 0 depending on whether or not R is supposed to be true in the interpretation of the sequence $o_1, \ldots, o_n$ of objects. We shall sometimes indicate what characteristic function a certain interpretation assigns to a predicate letter by stating a necessary and sufficient condition for the predicate letter to be true of a sequence of objects in the domain of the interpretation.

An *interpretation* is, of course, something that is an interpretation of some language. Our definition of 'interpretation of a language' should be taken as implying that each interpretation is the interpretation of only one language; so any interpretation assigns objects of the appropriate sort to at most enumerably many non-logical symbols. An interpretation of a *sentence* is, as one might guess, an interpretation of a language in which that sentence is a sentence; an interpretation of a *set of sentences* is an interpretation of all the sentences in the set.

A particularly important interpretation is *the standard interpretation of the language of arithmetic*, or '$\mathcal{N}$', for short. The domain of $\mathcal{N}$ is the set $\{0, 1, 2, \ldots\}$ of all natural numbers, and $\mathcal{N}$ assigns an appropriate sort of object to each of the symbols in L, the language of arithmetic. In fact, $\mathcal{N}$ assigns them just what you would expect (which is why it is the 'standard' interpretation): To $\mathbf{o}$ $\mathcal{N}$ assigns 0, the least natural number; to $'$ $\mathcal{N}$ assigns the successor function, whose value for each natural number n is $n+1$, the successor of n; to $+$ $\mathcal{N}$ assigns the addition function, whose value for any natural numbers m, n is $m+n$; and to $\cdot$ $\mathcal{N}$ assigns the multiplication function, whose value for any natural numbers is $m \cdot n$.

Example 9.1. An interpretation of the sentence (9.4) '*aLf*(*f*(*a*))'

(1) *Domain*: the set {Alma, Max, Dan}.

(2) *Characteristic function* ϕ *of* '*L*': see the table.

		2nd argument		
	ϕ	Alma	Max	Dan
1st argument	Alma	1	1	1
	Max	1	0	1
	Dan	0	0	0

(3) *Denotation of* '*a*': Alma.

(4) *Function for* '*f*': ƒ (Alma) = Max, ƒ (Max) = Dan, ƒ (Dan) = Dan. Clearly '*f*' can't be read as *the father of* in this interpretation: Dan can't be his own father. The facts of life being what they are, '*f*' can only be read in that way when the domain is infinite. Intuitively, sentence (9.4) is true in this interpretation: ƒ(ƒ(Alma)) = ƒ(Max) = Dan, and indeed ϕ (Alma, Dan) = 1 – Alma does love Dan in this interpretation.

Example 9.2. An interpretation of sentence (9.2)

(1) *Domain*: {Alma, Bert, Clara}.

(2) *Characteristic function* ϕ *of* '*L*': see the table.

ϕ	Alma	Bert	Clara
Alma	1	0	0
Bert	1	0	0
Clara	0	1	0

(3) *Denotation of* '*a*': Alma.

Intuitively, sentence (9.2) is false in this interpretation, for since ϕ (Alma, Alma) = 1, Alma is one of her own lovers; loving herself, she loves one of her lovers; and so, one of Alma's lovers' lovers does love her, contrary to what (9.2) asserts.

Example 9.3. An interpretation of sentence (9.5)

(1) *Domain*: {0, 1, 2, ...}.

(2) *Function assigned to* ′: the successor function.

(3) *Function assigned to* +: the *multiplication* function.

(4) Denotation of **0**: zero.

Intuitively, sentence (9.5) is false in this interpretation, for according to the interpretation it asserts that two *times* three is five. (9.5) is true in $\mathscr{N}$, however, for according to $\mathscr{N}$ it says that two plus three is five.

Now let's make the notion of truth in an interpretation explicit and independent of intuition. If S is any sentence, and $\mathscr{I}$ is one of its interpretations, we want to define $\mathscr{I}(S)$, where

$$\mathscr{I}(S) = \begin{cases} 1 & \text{if } S \text{ is true in interpretation } \mathscr{I}, \\ 0 & \text{if } S \text{ is false in interpretation } \mathscr{I}. \end{cases}$$

We'll do this by successively reducing the question, 'What value does $\mathscr{I}$ assign to S?' to similar questions about shorter sentences and about interpretations closely related to $\mathscr{I}$, until finally we are concerned only with atomic sentences, for which such questions are answered separately.

A sentence that isn't atomic must have one of the seven forms:

$$-S,\ (S_1 \& S_2),\ (S_1 \vee S_2),\ (S_1 \rightarrow S_2),\ (S_1 \leftrightarrow S_2),\ \forall vF, \exists vF,$$

where the Ss are sentences, v is a variable, and F is a formula that contains no free occurrences of any variable other than v. (Remember that a sentence is a formula in which no variable has free occurrences; if $S = \forall vF$, the only variable that could possibly have free occurrences in F is v.) Let's consider these cases in turn.

Case 1. $\mathscr{I}(-S)=1$ if $\mathscr{I}(S)=0$;
$\mathscr{I}(-S)=0$ if $\mathscr{I}(S)=1$.

Case 2. $\mathscr{I}(S_1 \& S_2)=1$ if $\mathscr{I}(S_1)=\mathscr{I}(S_2)=1$;
$\mathscr{I}(S_1 \& S_2)=0$ if either $\mathscr{I}(S_1)=0$ or $\mathscr{I}(S_2)=0$ or both.

Case 3. $\mathscr{I}(S_1 \vee S_2)=1$ if either $\mathscr{I}(S_1)=1$ or $\mathscr{I}(S_2)=1$ or both;
$\mathscr{I}(S_1 \vee S_2)=0$ if $\mathscr{I}(S_1)=\mathscr{I}(S_2)=0$.

Case 4. $\mathscr{I}(S_1 \rightarrow S_2)=1$ if either $\mathscr{I}(S_1)=0$ or $\mathscr{I}(S_2)=1$ or both;
$\mathscr{I}(S_1 \rightarrow S_2)=0$ if $\mathscr{I}(S_1)=1$ and $\mathscr{I}(S_2)=0$.

Case 5. $\mathscr{I}(S_1 \leftrightarrow S_2)=1$ if $\mathscr{I}(S_1)=\mathscr{I}(S_2)$;
$\mathscr{I}(S_1 \leftrightarrow S_2)=0$ if $\mathscr{I}(S_1) \neq (S_2)$.

Here we have simply described the truth tables for the connectives.

Case 6. $\mathscr{I}(\forall v F) = 1$ if $\mathscr{I}^a_o(F_{\imath}, a) = 1$ for every o in the domain of $\mathscr{I}$;
$\mathscr{I}(\forall v F) = 0$ if $\mathscr{I}^a_o(F_v a) = 0$ for even one o in that domain.

Explanation. $F_v a$ is the sentence obtained by writing a in place of all free occurrences of v in F; a is required to be a name that does not occur in F (it does not matter which), and $\mathscr{I}^a_o$ is the interpretation which is just like $\mathscr{I}$ except that in it, the name a is assigned the designation o. $\mathscr{I}^a_o$ therefore always has the same domain as $\mathscr{I}$. ($\mathscr{I}$ may or may not assign a any designation at all; if by chance $\mathscr{I}$ already assigns a the designation o, then $\mathscr{I}^a_o = \mathscr{I}$; $\mathscr{I}^a_o$ is 'not defined' if o is not in the domain of $\mathscr{I}$.)

Case 7. $\mathscr{I}(\exists v F) = 1$ if $\mathscr{I}^a_o(F_v a) = 1$ for even one o in the domain of $\mathscr{I}$;
$\mathscr{I}(\exists v F) = 0$ if $\mathscr{I}^a_o(F_v a) = 0$ for every o in that domain.

Note. Again, a must not occur in F, and $\mathscr{I}^a_o$ must assign designation o to a, and otherwise be just like $\mathscr{I}$.

This enumeration of cases may dazzle the eye, but all that's being done, here, is to restate the familiar interpretations of the connectives and quantifiers in a form that allows us to write out reasonably neat-looking calculations.

If a sentence has none of the foregoing seven forms, it must be atomic, i.e. must either be a sentence letter or consist of the equals-sign flanked by a pair of *terms* (case 8) or consist of an n-place predicate letter followed by a string of n *terms* (case 9).

What is a term? A term is either a name or something obtained by writing n (shorter) terms in the blanks of an n-place function symbol.

Examples. a, $f(a)$, $f(f(a))$, $g(a, f(a))$, $g(g(f(f(a)), a), f(a))$. (Here we assume that f is a one-place, and g a two-place, function symbol.)

If t is a term, what's $\mathscr{I}(t)$, the denotation of t in interpretation $\mathscr{I}$? If t is a name, $\mathscr{I}(t)$ is given in part (2) of the definition of interpretation. Otherwise t is of the form $f(t_1, \ldots, t_n)$ and $\mathscr{I}(t) = \mathrm{f}(\mathscr{I}(t_1), \ldots, \mathscr{I}(t_n))$, where f is the function that $\mathscr{I}$ assigns to f – see part (3).

If the sentence S is a sentence letter, then $\mathscr{I}(S)$ is explicitly given in part (4) of the definition of interpretation. Otherwise we have

Case 8. $\mathscr{I}(t_1 = t_2) = 1$ if $\mathscr{I}(t_1) = \mathscr{I}(t_2)$;
$\mathscr{I}(t_1 = t_2) = 0$ if $\mathscr{I}(t_1) \neq \mathscr{I}(t_2)$.

Case 9. $\mathscr{I}(Rt_1 \ldots t_n) = \phi(\mathscr{I}(t_1), \ldots, \mathscr{I}(t_n))$, where ϕ is the characteristic function of the n-place predicate letter R in interpretation $\mathscr{I}$, and the ts are terms whose denotations are $\mathscr{I}(t_1), \ldots, \mathscr{I}(t_n)$ in interpretation $\mathscr{I}$.

Now let's see how this gadgetry works on our three examples.

Example 9.1 again

With $\mathscr{I}$ as in Example 9.1 we can evaluate the sentence '$aLffa$'. (Some parentheses omitted.)

$$\begin{aligned}
\mathscr{I}(aLffa) &= \phi(\mathscr{I}a, \mathscr{I}ffa) && \text{by case 9,}\\
&= \phi(\mathscr{I}a, \mathrm{f}\mathscr{I}fa) && \text{by the definition of } \mathscr{I}(t),\\
&= \phi(\mathscr{I}a, \mathrm{f}\mathrm{f}\mathscr{I}a) && \text{by the definition of } \mathscr{I}(t) \text{ again,}\\
&= \phi(\text{Alma}, \mathrm{f}\mathrm{f}\,\text{Alma}) && \text{since } \mathscr{I}a = \text{Alma,}\\
&= \phi(\text{Alma}, \mathrm{f}\,\text{Max}) && \text{since } \mathrm{f}\,\text{Alma} = \text{Max,}\\
&= \phi(\text{Alma}, \text{Dan}) && \text{since } \mathrm{f}\,\text{Max} = \text{Dan,}\\
&= 1 && \text{by the table for } \phi.
\end{aligned}$$

Example 9.2 again

With $\mathscr{I}$ as in Example 9.2 we can show that sentence (9.2),

$$\forall x\,[\exists y\,(xLy \,\&\, yLa) \rightarrow -xLa],$$

is false in $\mathscr{I}$: $\mathscr{I}(9.2) = 0$ if $\mathscr{I}^b_o(\exists y\,(bLy \& yLa) \rightarrow -bLa) = 0$ for $o =$ Alma or $o =$ Bert or $o =$ Clara. (Case 6, with $v =$ 'x' and using 'b' instead of 'a' since 'a' occurs in 9.2.) Let's try $o =$ Alma:
$\mathscr{I}^b_{\text{Alma}}(\exists y\,(bLy \,\&\, yLa) \rightarrow -bLa) = 0$ if (see case 4) we have both of these:

(i) $\mathscr{I}^b_{\text{Alma}}(\exists y\,(bLy \,\&\, yLa)) = 1$. By case 7, (i) holds iff

$$\mathscr{I}^{b}_{\text{Alma}}{}^{c}_{o}(bLc \,\&\, cLa) = 1$$

for even one o, e.g., $o =$ Alma: $\mathscr{I}^{b}_{\text{Alma}}{}^{c}_{\text{Alma}}(bLc \,\&\, cLa) = 1$ iff (by cases 2 and 9) ϕ(Alma, Alma) $= \phi$(Alma, Alma) $= 1$, which is true, by the table for ϕ.

(ii) $\mathscr{I}^b_{\text{Alma}}(-bLa) = 0$. By case 1, (ii) holds iff $\mathscr{I}^b_{\text{Alma}}(bLa) = 1$, and by case 9 that holds iff ϕ (Alma, Alma) $= 1$, which is true by the table for ϕ.

Then conditions (i) and (ii) both hold, and we have shown that $\mathscr{I}(9.2) = 0$.

Example 9.3 again

With $\mathscr{I}$ as in Example 9.3 we can show that sentence (9.5) is false in $\mathscr{I}$: Since $\mathscr{I}(\mathbf{0}) = 0$, $\mathscr{I}(\mathbf{0}') = 0 + 1 = 1$, $\mathscr{I}(\mathbf{0}'') = 1 + 1 = 2$,

$$\mathscr{I}(\mathbf{0}''') = 2 + 1 = 3, \quad \mathscr{I}(\mathbf{0}'''') = 3 + 1 = 4, \quad \text{and}$$

$$\mathscr{I}(\mathbf{0}''''') = 4 + 1 = 5.$$

But then

$$\mathscr{I}(\mathbf{0}''+\mathbf{0}''') = \mathscr{I}(\mathbf{0}'')\cdot\mathscr{I}(\mathbf{0}''') = 2\cdot 3 = 6,$$

and so $\mathscr{I}(\mathbf{0}''+\mathbf{0}''') \neq \mathscr{I}(\mathbf{0}'''''')$, from which it follows by case 8, that $\mathscr{I}(\mathbf{0}''+\mathbf{0}''' = \mathbf{0}'''''') = 0$.

In terms of the basic concept $\mathscr{I}(S)$ of the truth value (1 or 0) of a sentence S in one of its interpretations $\mathscr{I}$ we can define:

Satisfaction. $\mathscr{I}$ satisfies S (or, equivalently, *S is true in $\mathscr{I}$*) iff $\mathscr{I}(S) = 1$.

Satisfiability. S is *satisfiable* iff $\mathscr{I}(S) = 1$ for some $\mathscr{I}$.

Validity. S is *valid* iff $\mathscr{I}(S) = 1$ for every interpretation $\mathscr{I}$ of S.

Example 9.4. Sentence (9.1), '$\forall x(xLx \rightarrow \exists yxLy)$', is valid

To prove this, we deduce a contradiction from the assumption that

(i) $\mathscr{I}(\forall x(xLx \to \exists y\, xLy)) = 0$ and thus prove that there can be no such interpretation $\mathscr{I}$ of S. By case 6, (i) holds iff

(ii) $\mathscr{I}^a_o(aLa \to \exists y\, aLy) = 0$ for some o in the domain D of $\mathscr{I}$, and by case 4, (ii) holds iff we have both (iii) and (iv):

(iii) $\mathscr{I}^a_o(aLa) = 1$, i.e. by case 9, $\phi(o,o) = 1$.

(iv) $\mathscr{I}^a_o(\exists y\, aLy) = 0$, i.e. by case 7, $\mathscr{I}^{a\,b}_{o\,p}(aLb) = 0$ for each p in D, i.e. by case 9, $\phi(o,p) = 0$ for each p in D. Then in particular, with $p = o$, we must have $\phi(o,o) = 0$.

Then (iii) and (iv) contradict each other, and since (i) holds if and only if both (iii) and (iv) do, (i) is refuted, and (9.1) is seen to be valid.

We write '$\vdash S$' to indicate that S is valid, and write '$S_1 \vdash S_2$' to indicate that S_1 implies S_2, i.e., the inference from S_1 as premise to S_2 as conclusion is valid.

Implication: $S_1 \vdash S_2$ iff for every $\mathscr{I}$ which is an interpretation of both S_1 and S_2, $\mathscr{I}(S_2) = 1$ if $\mathscr{I}(S_1) = 1$.

It is easy to see that we have

$$S_1 \vdash S_2 \text{ iff } \vdash (S_1 \to S_2). \tag{9.7}$$

Proof. By the definition of validity and case 4, the second of these conditions holds iff $\mathscr{I}(S_2) = 1$ if $\mathscr{I}(S_1) = 1$ for every interpretation $\mathscr{I}$ of S_1 and S_2, and by definition of implication, this is precisely when the first condition holds.

Example 9.5. $a = b \vdash f(a) = f(b)$

To prove this we derive a contradiction from the assumption that $\mathscr{I}$ is an interpretation of both sentences for which we have both (i) and (ii):

(i) $\mathscr{I}(a = b) = 1$, i.e. by case 8, $\mathscr{I}(a) = \mathscr{I}(b) = o$, say.

(ii) $\mathscr{I}(f(a) = f(b)) = 0$, i.e. $\mathscr{I}(f(a)) \neq \mathscr{I}(f(b))$, i.e. by the definition of $\mathscr{I}(t)$, $\mathfrak{f}(\mathscr{I}(a)) \neq \mathfrak{f}(\mathscr{I}(b))$.

But by (i), $\mathscr{I}(a) = \mathscr{I}(b) = o$, and therefore by (ii), $\mathfrak{f}(o) \neq \mathfrak{f}(o)$, which is impossible since $\mathfrak{f}$, being a function, is single-valued.

We can generalize the definition of *implication* as follows:

$S_1, \ldots, S_n \vdash S_{n+1}$ iff for every $\mathscr{I}$ which is an interpretation of all $n+1$ of the Ss we have:

$$\text{If } \mathscr{I}(S_1) = \ldots = \mathscr{I}(S_n) = 1 \text{ then } \mathscr{I}(S_{n+1}) = 1. \tag{9.8}$$

The relationship (9.7) between implication and validity can then be generalized:

$$S_1, \ldots, S_n \vdash S_{n+1} \text{ iff } \vdash [(S_1 \,\&\, \ldots \,\&\, S_n) \to S_{n+1}]. \tag{9.9}$$

Proof. By (9.8), the first condition fails iff for some $\mathscr{I}$ we have $\mathscr{I}(S_1) = \ldots = \mathscr{I}(S_n) = 1$ and $\mathscr{I}(S_{n+1}) = 0$; and by the definition of validity and cases 2 and 4, this is precisely when the second condition fails.

A *model* of a sentence is an interpretation of that sentence which satisfies it. Then the following are different ways of saying the same thing:

$$\mathscr{I}(S) = 1,\ S \text{ is true in } \mathscr{I},\ \mathscr{I} \text{ satisfies } S,\ \mathscr{I} \text{ is a model of } S.$$

We'll also speak of interpretations as satisfying *sets* of sentences and as being models of sets of sentences. To make such talk quite general, we'll want to allow the case in which the set is empty. We'll use capital Greek gamma (Γ) and delta (Δ) to stand for sets of sentences, and we use the usual symbol ($\varnothing$) for the empty set. The union $\Gamma \cup \Delta$ of sets Γ and Δ is the set whose members are the members of Γ together with the members of Δ. The set whose only member is the sentence S is $\{S\}$; the set whose members are the sentences $S_1, \ldots, S_n$ is $\{S_1, \ldots, S_n\}$. We define:

Satisfaction. $\mathscr{I}$ satisfies Γ(= *$\mathscr{I}$ is a model of* Γ) iff $\mathscr{I}$ is a model of every sentence in Γ.

Implication. $\Gamma \vdash S$ iff $\mathscr{I}(S) = 1$ whenever $\mathscr{I}$ is a model of Γ and an interpretation of S. (We may read '$\Gamma \vdash S$' as 'S follows from Γ', 'S is a (logical) consequence of Γ', or 'Γ implies S'.)

Satisfiability. Γ is satisfiable iff it has a model.

These definitions are contrived so as to make the empty set satisfiable: $\varnothing$ is satisfiable iff it has a model, and according to the definition of *$\mathscr{I}$ is a model of Γ* (definition of *satisfaction*, above) $\varnothing$ has a model iff for for some $\mathscr{I}$, $\mathscr{I}$ satisfies $\varnothing$, which in turn means that *for all S, if S is in $\varnothing$ then* $\mathscr{I}(S) = 1$. This quantified conditional is true because the antecedent is false for every S. Therefore,

$$\varnothing \text{ is satisfiable.} \tag{9.10}$$

Indeed, $\varnothing$ is satisfied by every interpretation. Things go most smoothly if we contrive the definition of satisfaction so as to make (9.10) true; and we have done so.

Note that

$$\vdash S \text{ iff } \varnothing \vdash S. \tag{9.11}$$

Proof. The right-hand side holds iff $\mathscr{I}$ is a model of S whenever it is a model of $\varnothing$ and an interpretation of S. By (9.10) this comes to the same thing as: *every interpretation of S is a model of S*. And by definition of validity, that comes to the same thing as $\vdash S$.

Further, note the following important connection between the concepts of implication and satisfiability:

$$\Gamma \vdash S \text{ iff } \Gamma \cup \{-S\} \text{ is unsatisfiable.} \tag{9.12}$$

Proof. $\Gamma \vdash S$ iff every interpretation of S which is a model of Γ is a model of S. By quantificational logic, this statement is equivalent to: for no $\mathscr{I}$ is $\mathscr{I}$ an interpretation of S *and* a model of Γ *and* not a model of S. But an interpetation of S which is not a model of S assigns value 0 to S and thus (case 1) assigns value 1 to $-S$; such an interpretation of S is a model of $-S$. Then the statement is equivalent to this: no $\mathscr{I}$ is a model of Γ *and* of $-S$, and by definition of *satisfiability*, *that* statement is equivalent to the claim that $\Gamma \cup \{-S\}$ is unsatisfiable.

In later chapters, we shall make use of the notion of a *theory* over and over again. This is the appropriate place to define it. A *theory* is a set whose members are just the sentences in some language that follow from the set. So if T is a theory, then for some language K, all members of T are sentences of K, and any sentence of K that follows from T is also a member of T. As '$\forall x\, x = x$' is a sentence of every language, and one that follows from any set whatsoever, every theory must contain '$\forall x\, x = x$'; thus the empty set is not a theory. The members of a theory T are ordinarily called its *theorems*, and we write: $\vdash_T A$ to mean that A is a theorem of T.

Q is a theory whose acquaintance we shall make in Chapter 14: it is the set of consequences in the language of arithmetic of the set

$$\begin{aligned}
&\{\forall x \forall y (x' = y' \rightarrow x = y),\\
&\forall x\, \mathbf{0} \neq x',\\
&\forall x (x \neq \mathbf{0} \rightarrow \exists y\, x = y'),\\
&\forall x\, x + \mathbf{0} = x,\\
&\forall x \forall y\, x + y' = (x+y)',\\
&\forall x\, x \cdot \mathbf{0} = \mathbf{0},\\
&\forall x \forall y\, x \cdot y' = (x \cdot y) + x\}.
\end{aligned}$$

All of the members of this set are true in $\mathscr{N}$, and hence all the theorems of Q are true in $\mathscr{N}$. As we shall see later, the converse is definitely not the case! There are sentences of L that are true in $\mathscr{N}$, but not consequences of Q. (In fact, '$\forall x \forall y\, x+y = y+x$' is one of them; see Exercise 14.2.)

Our definition of a theory T is quite general: we do not require that any sentences in T be singled out as 'axioms', we do not require that there be a finite or even an effectively specifiable set of sentences in T from which all the others follow, we do not require that there be any effective procedure for deciding whether any given sentence is a theorem of the theory, and we do not require that a theory be satisfiable. All that is required of a set of sentences T for it to be a theory is that there be a language with respect to which T is closed under logical consequence, i.e. that any consequence of the theory that is in the language also be a member of the theory.

The last definition in this series is

Logical equivalence. $S_1 \simeq S_2$ (S_1 is logically equivalent to S_2) iff for all $\mathscr{I}$, if $\mathscr{I}$ is an interpretation of S_1 and of S_2, then $\mathscr{I}(S_1) = \mathscr{I}(S_2)$.

It should be obvious that $\simeq$ is an *equivalence relation on the set of sentences* in the technical sense that $\simeq$ is:

Reflexive on the set of sentences: for all sentences S, $S \simeq S$.
Symmetrical. For all S_1 and S_2, if $S_1 \simeq S_2$, then $S_2 \simeq S_1$.
Transitive. For all S_1, S_2, S_3, if $S_1 \simeq S_2$ and $S_2 \simeq S_3$, then $S_1 \simeq S_3$.

Logical implication and equivalence are related as follows:

$$S_1 \simeq S_2 \quad \text{iff} \quad S_1 \vdash S_2 \quad \text{and} \quad S_2 \vdash S_1. \tag{9.13}$$

Proof. The definition of implication can be reformulated as: $S \vdash T$ iff for every $\mathscr{I}$ which is an interpretation both of S and T, $\mathscr{I}(S) \leqslant \mathscr{I}(T)$. Since we have both $\mathscr{I}(S_1) \leqslant \mathscr{I}(S_2)$ and $\mathscr{I}(S_2) \leqslant \mathscr{I}(S_1)$ iff $\mathscr{I}(S_1) = \mathscr{I}(S_2)$, (9.13) then follows from the definition of $\simeq$.

It will be useful to generalize our definitions of the relations $\vdash$ and $\simeq$ so that they can be asserted to hold between formulas that aren't sentences, e.g.

$$(\exists x\, Px \to Py), \quad \text{and} \quad \forall x (Px \to Py),$$

where the variable y has a free occurrence in each formula. In particular, it would be good to have a sense of $\simeq$ in which it would be true that

> *when logical equivalents are substituted for each other, the results are logically equivalent*; i.e. if F_1 is a subformula of G_1, and F_2 is substituted for an occurrence of F_1 in G_1 to get a new formula G_2, then $G_2 \simeq G_1$ if $F_2 \simeq F_1$. (9.14)

Thus the logical equivalence noted above would allow us to conclude that

$$\underbrace{\exists y \overbrace{(\exists x\, Px \to Py)}^{F_1}}_{G_1} \simeq \underbrace{\exists y \overbrace{\forall x (Px \to Py)}^{F_2}}_{G_2}$$

The following generalized definitions do the job:

Implication. $F_1 \vdash F_2$ iff $\mathscr{I}(F_1^*) \leqslant \mathscr{I}(F_2^*)$ for every $\mathscr{I}$ which is an interpretation of both F_1^* and F_2^*. Here, F_1^* and F_2^* are the results of substituting names $a_1, \ldots, a_n$ for all variables $v_1, \ldots, v_n$ that have free occurrences in F_1 or F_2, with distinct names being substituted for distinct variables, and with all of $a_1, \ldots, a_n$ distinct from all names that occur in F_1 or F_2. Where a variable occurs free in both F_1 and F_2, the same name is to be substituted for it in both formulas.

Logical equivalence. $F_1 \simeq F_2$ iff $\mathscr{I}(F_1^*) = \mathscr{I}(F_2^*)$ for every $\mathscr{I}$ which is an interpretation of both F_1^* and F_2^*.

Example 9.6. $F_1 \simeq F_2$, where $F_1 = (\exists x Px \to Py)$, $F_2 = \forall x(Px \to Py)$

The claim is that $\mathscr{I}(F_1^*) = \mathscr{I}(F_2^*)$ for each $\mathscr{I}$ for which both sides of the equation are defined, where we may take $F_1^* = (\exists x\, Px \to Pa)$ and $F_2^* = \forall x (Px \to Pa)$. To prove the claim, consider the two sides separately:

$\mathscr{I}(F_1^*) = 1$ iff $\mathscr{I}(\exists x\, Px) = 0$ or $\mathscr{I}(Pa) = 1$ (case 4)
iff either for each o $\mathscr{I}_o^b(Pb) = 0$ or $\mathscr{I}(Pa) = 1$ (case 7).
$\mathscr{I}(F_2^*) = 1$ iff for each o, either $\mathscr{I}_o^b(Pb) = 0$ or $\mathscr{I}_o^b(Pa) = 1$ (cases 6, 4).

As the names a and b are distinct, $\mathscr{I}_o^b(Pa) = \mathscr{I}(Pa)$ for each o, and so the two conditions are equivalent.

It should be fairly clear, now, why (9.14) holds. Thus, with F_1 and F_2 as above and $G_1 = \exists y\, F_1$ and $G_2 = \exists y\, F_2$, we have $G_1 \simeq G_2$ because in applying case 7 to $\mathscr{I}(G_1)$ and $\mathscr{I}(G_2)$, $F_{1\,v}a$ and $F_{2\,v}a$ are respectively F_1^* and F_2^*. It should also be evident that (9.13) holds with formulas F_1 and F_2 in place of sentences S_1 and S_2.

Before we conclude this chapter, we shall use (9.14) to prove that any formula can be put into *prenex normal form*. A formula is in *prenex form* iff all quantifiers (if any) occur at the extreme left, without intervening parentheses: The form is

$$Q_1 v_1 \ldots Q_n v_n F,$$

where F is a quantifier-free formula, and each Q_i is either $\forall$ or $\exists$. Thus in Example 9.6, F_2 is in prenex form but F_1 is not; and since F_2 is a prenex formula logically equivalent to F_1, F_2 is a prenex form of F_1. Then the claim is,

> Corresponding to each formula F_1 there is a formula F_2 where F_2 is prenex and $F_2 \simeq F_1$. (9.15)

To prove (9.15) it is sufficient to note that the following are really logical equivalences, and that with their use one can successively move quantifiers to the left (in a sequence of logically equivalent formulas) until finally a prenex formula is obtained.

$$-\forall v\, F \simeq \exists v - F; \quad -\exists v\, F \simeq \forall v - F. \qquad (9.16)$$

Here the general form is

$$-Qv\, F \simeq Q'v - F,$$

where Q' is $\exists$ if Q is $\forall$, and Q' is $\forall$ if Q is $\exists$.

Provided that v does not occur free in G,

$$\left.\begin{array}{ll} (a) & (Qv\, F \,\&\, G) \simeq Qv(F \,\&\, G), \\ (b) & (G \,\&\, Qv\, F) \simeq Qv(G \,\&\, F), \\ (c) & (Qv\, F \vee G) \simeq Qv(F \vee G), \\ (d) & (G \vee Qv\, F) \simeq Qv(G \vee F), \\ (e) & (G \to Qv\, F) \simeq Qv(G \to F), \\ (f) & (Qv\, F \to G) \simeq Q'v(F \to G). \end{array}\right\} \qquad (9.17)$$

$$Qv\, F \simeq Qw\, F_v w, \qquad (9.18)$$

where w does not occur in F and $F_v w$ is the result of replacing all free occurrences of the variable v in F by occurrences of the variable w.

Example 9.7. A prenex normal form of $(Qv\,F \leftrightarrow G)$, where v does not occur free in G, and F and G contain no quantifiers

We first put the biconditional into a logically equivalent form,

$$(Qv\,F \rightarrow G)\,\&\,(G \rightarrow Qv\,F).$$

By (9.17*e*) and (9.17*f*) this becomes $Q'v(F \rightarrow G)$ & $Qv(G \rightarrow F)$, and by (9.17*a*) it becomes $Q'v((F \rightarrow G)\,\&\,Qv(G \rightarrow F))$. Now v will presumably occur free in the first conjunct $(F \rightarrow G)$, so we must use (9.18) to get $Q'v((F \rightarrow G)\,\&\,Qw(G \rightarrow F_v w))$, where w is new to the entire formula. We can then apply (9.17*b*) to get the prenex form

$$Q'v\,Qw((F \rightarrow G)\,\&\,(G \rightarrow F_v w)).$$

If we had brought the quantifiers out in the other order, a different (but logically equivalent) prenex form would have been obtained.

In general, biconditionals must be replaced in formulas containing them by appropriate equivalents (such as conjunctions of suitable conditionals) before the prenexing operations can be applied.

The following further equivalences are sometimes useful in obtaining prenex forms with as few quantifiers as possible.

$$Qv\,F \simeq F, \text{ if } v \text{ does not occur free in } F. \qquad (9.19)$$

$$(\forall v\,F\,\&\,\forall v\,G) \simeq \forall v(F\,\&\,G). \qquad (9.20)$$

$$(\exists v\,F \vee \exists v\,G) \simeq \exists v(F \vee G). \qquad (9.21)$$

$$(\forall v\,F \rightarrow \exists v\,G) \simeq \exists v(F \rightarrow G). \qquad (9.22)$$

Exercises

9.1 Equivalences (9.20)–(9.22) fail when the quantifiers are all changed: in each case, implication then holds in one direction but not in the other. Prove that

$$(\exists v\,F\,\&\,\exists v\,G) \nvdash \exists v(F\,\&\,G), \quad \text{but} \quad \exists v(F\,\&\,G) \vdash (\exists v\,F\,\&\,\exists v\,G). \qquad (9.23)$$

$$(\forall v\,F \vee \forall v\,G) \vdash \forall v(F \vee G), \quad \text{but} \quad \forall v(F \vee G) \nvdash (\forall v\,F \vee \forall v\,G). \qquad (9.24)$$

$$(\exists v\,F \rightarrow \forall v\,G) \vdash \forall v(F \rightarrow G), \quad \text{but} \quad \forall v(F \rightarrow G) \nvdash (\exists v\,F \rightarrow \forall v\,G). \qquad (9.25)$$

9.2 Test for validity and put into prenex form

(*a*) $\exists x(Px \rightarrow \forall x\,Px)$,

(*b*) $\exists x(\exists x\,Px \rightarrow Px)$.

9.3 Put into prenex form

$$\{\forall x[Sx \rightarrow \exists y(Py\,\&\,yOx)] \leftrightarrow \exists x[Px\,\&\,\forall y(Sy \rightarrow xOy)]\}.$$

Partial solutions

9.1 Counterexamples to the invalid ones, i.e. interpretations that make premise true and conclusion false: In each case the domain is $\{o_1, o_2\}$; characteristic functions of F and G are ϕ_1 and ϕ_2, respectively. For (9.23) and (9.24), let $\phi_1(o_1) = \phi_2(o_2) = 0$ and $\phi_1(o_2) = \phi_2(o_1) = 1$. For (9.25) let $\phi_1(o_1) = \phi_2(o_1) = 0$ and $\phi_2(o_2) = \phi_1(o_2) = 1$.

9.2 (*a*) $\mathscr{I}(\exists x(Px \rightarrow \forall x\,Px)) = 0$ *iff* $\mathscr{I}^{a}_{o}(Pa \rightarrow \forall x\,Px) = 0$ for each o, *iff* for each o, $\phi(o) = 1$ and $\mathscr{I}^{a\,b}_{o\,p}(Pb) = 0$ for some p, *iff* for each o, $\phi(o) = 1$ and $\phi(p) = 0$ for some p. With $o = p$ this is a contradiction. Then (*a*) is valid. Prenex form: $\exists x\,\forall y(Px \rightarrow Py)$. (*b*) Valid. Prenex form:

$$\exists x\,\forall y(Py \rightarrow Px).$$

9.3 One answer is

$$\begin{aligned}&\exists x_1\forall y_1\exists x_2\forall y_2\forall x_3\exists y_3\forall x_4\exists y_4(\{[Sx_1 \rightarrow (Py_1\,\&\,y_1Ox_1)]\\ &\quad \rightarrow [Px_2\,\&\,(Sy_2 \rightarrow x_2Oy_2)]\}\,\&\,\{[Px_3\,\&\,(Sy_3 \rightarrow x_3Oy_3)]\\ &\quad \rightarrow [Sx_4 \rightarrow (Py_4\,\&\,y_4\,Ox_4)]\}).\end{aligned}$$

10
First-order logic is undecidable

In Chapters 11 and 12 we shall demonstrate the existence of a mechanical positive test for first-order unsatisfiability or, what comes to the same thing, a mechanical positive test for first-order validity. We shall now demonstrate the non-existence of any corresponding negative tests. We shall thereby have proved the unsolvability of the decision problem for first-order satisfiability and validity: we shall have proved first-order logic to be *undecidable.*

In general, the *decision problem* for a property is solvable if there is a mechanical test (= a computational routine, an effective procedure) which, applied to *any* object of the appropriate sort, *eventually* classifies that object *correctly* as a positive or a negative instance of that property. ('Eventually' here means 'after some finite number of steps'.) A positive test for a property is a mechanical test which eventually classifies as positive all and only its positive instances. A negative test is one which eventually classifies as negative all and only the negative instances. If both a positive and a negative test for a property exist, then, and only then, is the decision problem for that property solvable; for since any appropriate object will be either a positive or a negative instance, if one is equipped with both sorts of test, one can apply both to the object – dividing one's time so as to alternate steps of the two tests – and thus eventually discover which sort of instance the object is. (Conversely, any test which eventually classifies correctly any object as either a positive or a negative instance counts as both a positive and a negative test.) Here, the properties that interest us are validity and satisfiability, and the 'objects of the appropriate sort' are sentences in the notation of first-order logic. The derivations to be introduced in Chapter 11 will give us a positive test for validity of sentences (and therewith a negative test for satisfiability). In this chapter we prove that those tests cannot be supplemented – by a positive test for satisfiability or a negative test for validity – so as to solve the decision problem for satisfiability or validity of sentences of first-order logic.

Our proof that there is no solution to the decision problem for first-order validity will take the form of a *reductio ad absurdum*: we shall prove that if there were such a test, then the halting problem would be

solvable, i.e., there would be a computational routine for discovering whether or not Turing machines eventually halt, when started in state q_1 scanning the leftmost of a string of 1's on an otherwise blank tape. But at the very end of Chapter 5 we saw that the halting problem is unsolvable, if Church's thesis is correct. In particular, we proved that no Turing machine can carry out a computational routine which solves the halting problem, and we noted that by Church's thesis this means that there is no computational routine of any sort which solves the halting problem.

The *reductio* will go in this way: we show how, given the machine table or flow graph or other suitable description of a Turing machine, and any n, we can effectively write down a *finite* set Δ of sentences and a sentence H such that $\Delta \vdash H$ if and only if the machine in question does eventually halt when given input n, i.e. when started in state q_1 scanning the leftmost of an unbroken string of n 1s on an otherwise blank tape. For each machine and input, we also specify an interpretation $\mathscr{I}$. Under $\mathscr{I}$, the sentence H will say that the machine eventually halts, and the sentences in Δ will describe the operation of the machine, will say that its input is n, and will also say something about the successor function $'$ (where for any integer i, $i' = i+1$). Thus, if we could solve the decision problem for validity of sentences we could effectively determine whether or not the machine eventually halts, for we have $\Delta \vdash H$ if and only if a certain sentence is valid, *viz.*, the conditional whose antecedent is the conjunction of all the sentences in Δ and whose consequent is H.

The thing is really quite simple.† We shall have no need to suppose that only the symbols S_0 and S_1 are used, nor need we suppose that the tape is infinite only to the right. We imagine that the squares of the tape are numbered:

…	−3	−2	−1	0	1	2	3	…

We imagine that time is broken up into a series of moments t at which machines perform exactly one of their operations, and that there is a moment 0 at which our machine starts, scanning square 0. The moments of time are supposed to extend endlessly into the future and the past, just as the tape is supposed to extend endlessly to the right and the left.

† Simple, anyway, now that J. R. Büchi has shown us how to do it simply: see his 'Turing machines and the *Entscheidungsproblem*', *Math. Annalen* 148 (1962), 201–13. That there is no decision procedure for first-order validity was first shown by Alonzo Church.

We assume that the machine is 'plugged in' at moment o and 'unplugged' at the first moment (if any) after the one at which it halts; we assume that at all negative times and at all times later than the first one (if any) at which it halts, the machine is in none of its states, scanning none of its squares, and that no symbol (not even the blank) occurs anywhere on its tape.

For each state q_i which the machine can be in, we pick a two-place predicate letter Q_i, and for each symbol S_j which the machine can read or print, we pick a two-place predicate letter, which we shall also designate by 'S_j'. In addition to the Q_is and the S_js (and '$=$' and the usual logical apparatus) the only symbols that occur in the sentences in $\Delta \cup \{H\}$ are the name $\mathbf{o}$, the one-place function symbol $'$, and the two-place predicate letter $<$.

In the intended interpretation $\mathscr{I}$ of the sentences in $\Delta \cup \{H\}$, the variables range over the integers – positive, negative, and zero. $\mathscr{I}$ assigns zero to $\mathbf{o}$ and the successor function to $'$. The Q_is, the S_js and $<$ are interpreted as follows:

$\mathscr{I}$ stipulates that Q_i is to be true of t, x *iff* at time t, the machine is in state q_i, scanning square number x;

$\mathscr{I}$ stipulates that S_j is to be true of t, x *iff* at time t, the symbol S_j is in square number x; and

$\mathscr{I}$ stipulates that $<$ is to be true of x, y *iff* x is less than y.

We now say what the sentences of Δ are. (We shall use 't' as a variable when a time is intended, and 'x' and 'y' as variables when tape squares are intended, in order to remind the reader of the intended interpretation. Formally, the function of a variable is signalled by its position at the left (time) or right (tape square) of the symbols Q_i and S_j.) Suppose the machine can read or print the symbols $S_0, \ldots, S_r$. Corresponding to the three sorts of expression that can appear to the right of the colon over an arrow of a flow graph of the machine, we have three sorts of sentences which, under interpretation $\mathscr{I}$, describe features of the machine's operation.

For each label in the flow graph of form: $\textcircled{i} \xrightarrow{S_j:S_k} \textcircled{m}$ we have in Δ the sentence

$$\forall t \forall x \forall y \{[tQ_i x \,\&\, tS_j x] \rightarrow [t'Q_m x \,\&\, t'S_k x \,\&\, (y \neq x \\ \rightarrow (tS_0 y \rightarrow t'S_0 y) \,\&\, \ldots \,\&\, (tS_r y \rightarrow t'S_r y))]\}. \quad (10.1)$$

Here is an English translation of (10.1) under the intended interpretation $\mathscr{I}$:

If the machine is in state q_i at time t and is then scanning square number x on which symbol S_j occurs, then at time $t+1$ the machine is in state q_m scanning square number x, where the symbol S_k occurs, and in all squares other than x, the same symbols appear at time $t+1$ as appeared at time t (for all t and x).

The case is not excluded in which $i = m$; the diagram would then be: $\overset{S_j:S_k}{\circlearrowright}\textcircled{i}$ Corresponding remarks apply to the diagrams below.

For each label in the flow graph of form: $\textcircled{i} \xrightarrow{S_j:R} \textcircled{m}$ we have in Δ the sentence

$$\forall t \forall x \forall y \{[tQ_i x \,\&\, tS_j x] \\ \rightarrow [t'Q_m x' \,\&\, (tS_0 y \rightarrow t'S_0 y) \,\&\, \ldots \,\&\, (tS_r y \rightarrow t'S_r y)]\}. \quad (10.2)$$

And for each label in the graph of form: $\textcircled{i} \xrightarrow{S_j:L} \textcircled{m}$ we have in Δ the sentence

$$\forall t \forall x \forall y \{[tQ_i x' \,\&\, tS_j x'] \\ \rightarrow [t'Q_m x \,\&\, (tS_0 y \rightarrow t'S_0 y) \,\&\, \ldots \,\&\, (tS_r y \rightarrow t'S_r y)]\}. \quad (10.3)$$

One sentence in Δ says that initially the machine is in state q_1 scanning the leftmost of an unbroken string of n 1s on an otherwise blank tape:

$$\mathbf{o}Q_1\mathbf{o} \,\&\, \mathbf{o}S_1\mathbf{o} \,\&\, \mathbf{o}S_1\mathbf{o}' \,\&\, \ldots \,\&\, \mathbf{o}S_1\mathbf{o}^{(n-1)} \\ \&\, \forall y [(y \neq \mathbf{o} \,\&\, y \neq \mathbf{o}' \,\&\, \ldots \,\&\, y \neq \mathbf{o}^{(n-1)}) \rightarrow \mathbf{o}S_0 y]. \quad (10.4)$$

'$\mathbf{o}^{(n-1)}$' here abbreviates the result of attaching n successor symbols to the symbol $\mathbf{o}$. Note that if there are n 1s on the tape at time 0, the leftmost of them is in square 0, so the rightmost must be in square $n-1$, not square n. If $n = 0$, (10.4) is

$$\mathbf{o}Q_1\mathbf{o} \,\&\, \forall y\, \mathbf{o}S_0 y.$$

One sentence in Δ says that each integer is the successor of exactly one integer:

$$\forall z \exists x\, z = x' \,\&\, \forall z \forall x \forall y (z = x' \,\&\, z = y' \rightarrow x = y). \quad (10.5)$$

We require a guarantee that if p and q are different natural numbers, then the sentence $\forall x\, x^{(p)} \neq x^{(q)}$ is implied by Δ. All such sentences are consequences of the following:

$$\forall x \forall y \forall z (x < y \,\&\, y < z \rightarrow x < z) \,\&\, \forall x \forall y (x' = y \rightarrow x < y) \\ \&\, \forall x \forall y (x < y \rightarrow x \neq y). \quad (10.6)$$

Example: from (10.6) we can infer $x'' < x'''$ and $x''' < x''''$, whence $x'' < x''''$, whence $x'' \neq x''''$, i.e. $x^{(2)} \neq x^{(4)}$.

We take Δ to be the set of all the sentences (10.1), (10.2) and (10.3) which correspond to the various arrows in the machine's flow graph, together with the three additional sentences (10.4), (10.5) and (10.6). It is clear that $\mathscr{I}$ is a model of Δ.

As for H, we note that a machine halts at time t if it is then in a state q_i scanning a symbol S_j and there is no entry for q_i, S_j in its machine table. (Otherwise put: there is no arrow in the machine's flow chart leaving node i before whose colon S_j occurs.) So for our sentence H, we take the disjunction of all sentences

$$\exists t \exists x (tQ_i x \,\&\, tS_j x) \tag{10.7}$$

such that there is no entry for q_i, S_j in the table of our machine. If there is always an entry for every q_i, S_j, then the machine never halts and we take H to be some sentence that is false in $\mathscr{I}$, e.g. $\mathbf{o} \neq \mathbf{o}$.

And there we have it: given a machine and an input n, we have shown how to find a finite set Δ of sentences and a sentence H such that (so we claim) we have $\Delta \vdash H$ *if and only if the machine eventually halts when given n as an input*. Of course we now have to *verify* that claim. Our proof of this fact will appeal freely to various facts about first-order logical entailment (to facts about $\vdash$) with which the reader is presumed to be familiar.

They can be verified by the method of Chapter 11, or by more direct arguments in the terms of Chapter 9, or by use of other methods of proof (e.g. natural deduction or trees).

The verification

The 'only if' part is trivial. All of the sentences in Δ are true in the intended interpretation $\mathscr{I}$. Therefore if $\Delta \vdash H$, H is true in $\mathscr{I}$. But H is true in $\mathscr{I}$ if and only if the machine eventually halts with input n.

The 'if' part is harder. ([illegible] $\Delta \vdash H$)

First, we need a convention about 'negative numerals': if p is a negative integer, and $p = -q$, then the formulas

$$x Q_i \mathbf{o}^{(p)},$$
$$x S_j \mathbf{o}^{(p)},$$
$$y \neq \mathbf{o}^{(p)}$$

are to be regarded as abbreviations of the formulas

$$\exists z (x Q_i z \,\&\, z^{(q)} = \mathbf{o}),$$
$$\exists z (x S_j z \,\&\, z^{(q)} = \mathbf{o}),$$
$$\exists z (y \neq z \,\&\, z^{(q)} = \mathbf{o}),$$

respectively. (Similar formulas are to be disabbreviated in a similar way.)

We now introduce a special kind of sentence, called a *description of time s*. A description of time s is a sentence which says, in the obvious way, what state the machine is in at time s, what square it is then scanning, and what symbols are on what squares of the tape, and does so by using the language of the sentences in $\Delta \cup \{H\}$. More precisely, a sentence is a description of time s if it has this form:

$$\mathbf{o}^{(s)} Q_i \mathbf{o}^{(p)} \,\&\, \mathbf{o}^{(s)} S_{j_1} \mathbf{o}^{(p_1)} \,\&\, \ldots \,\&\, \mathbf{o}^{(s)} S_j \mathbf{o}^{(p)} \,\&\, \ldots \,\&\, \mathbf{o}^{(s)} S_{j_v} \mathbf{o}^{(p_r)} \,\&$$
$$\forall y [(y \neq \mathbf{o}^{(p_1)} \,\&\, \ldots \,\&\, y \neq \mathbf{o}^{(p)} \,\&\, \ldots \,\&\, y \neq \mathbf{o}^{(p_r)}) \to \mathbf{o}^{(s)} S_0 y]. \tag{10.8}$$

Here, we require that $p_1, \ldots, p, \ldots, p_r$ be an increasing sequence of integers; p *may* be p_1 or p_r. *Example*: (10.4) is a description of time 0.

Suppose now that the machine eventually halts with input n. Then for some s, i, p, and j, at time s the machine is in state q_i scanning square number p on which the symbol S_j occurs, but there is no entry for q_i, S_j in its machine table.

Suppose further that Δ implies some description G of time s. Since $\mathscr{I}$ is a model of Δ, G will be true in $\mathscr{I}$. Therefore two of the conjuncts of G will be $\mathbf{o}^{(s)} Q_i \mathbf{o}^{(p)}$ and $\mathbf{o}^{(s)} S_j \mathbf{o}^{(p)}$, and therefore G will imply

$$\exists t \exists x (tQ_i x \,\&\, tS_j x),$$

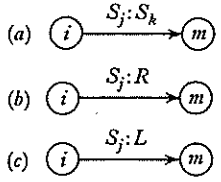
which is one of the disjuncts of H. Therefore Δ will imply H.

Therefore we need only show that for each s which is not negative, *if the machine has not halted before time s, then Δ implies some description of time s*. We prove this by mathematical induction on s.

Basis step. $s = 0$. Δ contains, and hence implies (10.4), which is a description of time 0.

Induction step. Suppose the italicized statement is true (for s). Suppose further that the machine has not halted before time $s+1$. Then the machine has not halted before time s and does not halt at time s. Then Δ implies some description (10.8) of time s. We must show that Δ implies some description of time $s+1$.

Since $\mathscr{I}$ is a model of Δ, (10.8) is true in $\mathscr{I}$. Therefore at time s, the machine is in state q_i, scanning some square (number p) on which the symbol S_j occurs. Since the machine does not halt at s, there must appear in its flow graph an arrow of one of the three forms



If (a) holds, then one of the sentences of Δ is

$$\forall t\,\forall x\,\forall y\,\{[tQ_ix\,\&\,tS_jx] \rightarrow [t'Q_mx\,\&\,t'S_kx$$
$$\&\,(y \neq x \rightarrow ((tS_0y \rightarrow t'S_0y)\,\&\,\ldots\,\&\,(tS_ry \rightarrow t'S_ry)))]\}.$$

This, together with (10.5), (10.6) and (10.8), implies

$$\mathbf{o}^{(s+1)}Q_m\mathbf{o}^{(p)}\,\&\,\mathbf{o}^{(s+1)}S_{j_1}\mathbf{o}^{(p_1)}\,\&\,\ldots\,\&\,\mathbf{o}^{(s+1)}S_k\mathbf{o}^{(p)}\,\&\,\ldots\,\&\,\mathbf{o}^{(s+1)}S_{j_v}\mathbf{o}^{(p_v)}$$
$$\&\,\forall y\,[(y \neq \mathbf{o}^{(p_1)}\,\&\,\ldots\,\&\,y \neq \mathbf{o}^{(p)}\,\&\,\ldots\,\&\,y \neq \mathbf{o}^{(p_r)}) \rightarrow \mathbf{o}^{(s+1)}S_0y]$$

which is a description of time $s+1$.

If (b) holds, then one of the sentences of Δ is

$$\forall t\,\forall x\,\forall y\,\{[tQ_ix\,\&\,tS_jx] \rightarrow [t'Q_mx'\,\&\,(tS_0y \rightarrow t'S_0y)\,\&\,\ldots$$
$$\&\,(tS_ry \rightarrow t'S_ry)]\}.$$

There is some symbol S_q such that this, together with (10.5), (10.6) and (10.8), implies

$$\mathbf{o}^{(s+1)}Q_m\mathbf{o}^{(p+1)}\,\&\,\mathbf{o}^{(s+1)}S_{j_1}\mathbf{o}^{(p_1)}\,\&\,\ldots\&\,\mathbf{o}^{(s+1)}S_j\mathbf{o}^{(p)}\,\&\,\mathbf{o}^{(s+1)}S_q\mathbf{o}^{(p+1)}\,\&\,\ldots$$
$$\&\,\mathbf{o}^{(s+1)}S_{j_v}\mathbf{o}^{(p_v)}\,\&\,\forall y[(y \neq \mathbf{o}^{(p_1)}\,\&\,\ldots$$
$$\&\,y \neq \mathbf{o}^{(p)}\,\&\,y \neq \mathbf{o}^{(p+1)}\,\&\,\ldots\,\&\,y \neq \mathbf{o}^{(p_r)}) \rightarrow \mathbf{o}^{(s+1)}S_0y]$$

which is a description of time $s+1$.

If (c) holds, then one of the sentences of Δ is

$$\forall t\,\forall x\,\forall y\,\{[tQ_ix'\,\&\,tS_jx'] \rightarrow [t'Q_mx\,\&\,(tS_0y \rightarrow t'S_0y)\,\&\,\ldots$$
$$\&\,(tS_ry \rightarrow t'S_ry)]\}.$$

There is some symbol S_q such that this, together with (10.5), (10.6) and (10.8), implies

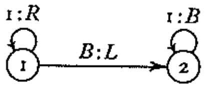
$$\mathbf{o}^{(s+1)}Q_m\mathbf{o}^{(p-1)}\,\&\,\mathbf{o}^{(s+1)}S_{j_1}\mathbf{o}^{(p_1)}\,\&\,\ldots\,\&\,\mathbf{o}^{(s+1)}S_q\mathbf{o}^{(p-1)}\,\&\,\mathbf{o}^{(s+1)}S_j\mathbf{o}^{(p)}\,\&\,\ldots$$
$$\&\,\mathbf{o}^{(s+1)}S_{j_v}\mathbf{o}^{(p_v)}\,\&\,\forall y\,[(y \neq \mathbf{o}^{(p_1)}\,\&\,\ldots$$
$$\&\,y \neq \mathbf{o}^{(p-1)}\,\&\,y \neq \mathbf{o}^{(p)}\,\&\,\ldots\,\&\,y \neq \mathbf{o}^{(p_r)}) \rightarrow \mathbf{o}^{(s+1)}S_0y]$$

which is a description of time $s+1$.

In all three cases Δ implies a description of time $s+1$ and therewith the undecidability of first-order logic is proved.

Exercises

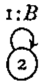
10.1 Write out the sentences in Δ and the sentence H for the following machine with $n = 2$, $n = 1$, and $n = 0$, and verify that $\Delta \vdash H$ in case $n = 0$.



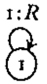
10.2 In each of cases (a), (b), and (c), verify that Δ implies the description of time $s+1$.

Solutions

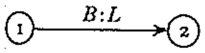
10.1 Members of Δ corresponding to the arrows:



$$tQ_2x\,\&\,tS_1x \rightarrow t'Q_2x\,\&\,t'S_0x\,\&\,(y \neq x \rightarrow (tS_0y \rightarrow t'S_0y)$$
$$\&\,(tS_1y \rightarrow t'S_1y)).$$



$$tQ_1x\,\&\,tS_1x \rightarrow t'Q_1x'\,\&\,(tS_0y \rightarrow t'S_0y)\,\&\,(tS_1y \rightarrow t'S_1y).$$



$$tQ_1x'\,\&\,tS_0x' \rightarrow t'Q_2x\,\&\,(tS_0y \rightarrow t'S_0y)\,\&\,(tS_1y \rightarrow t'S_1y).$$

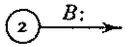
(Here, universal quantifiers ∀t, ∀x, ∀y, braces { } and brackets [ ] have been suppressed, following the convention that free variables are to be read as universally quantified and that '&' has more binding force than '→'.) The member of Δ which describes time 0 will be one thing or another, depending on the value of n.

If $n = 0$: $\mathbf{o}Q_1\mathbf{o}\,\&\,\forall y\,\mathbf{o}S_0y.$

If $n = 1$: $\mathbf{o}Q_1\mathbf{o}\,\&\,\mathbf{o}S_1\mathbf{o}\,\&\,\forall y(y \neq \mathbf{o} \rightarrow \mathbf{o}S_0y).$

If $n = 2$: $\mathbf{o}Q_1\mathbf{o}\,\&\,\mathbf{o}S_1\mathbf{o}\,\&\,\mathbf{o}S_1\mathbf{o}'\,\&\,\forall y(y \neq \mathbf{o}\,\&\,y \neq \mathbf{o}' \rightarrow \mathbf{o}S_0y).$

Finally, we have (10.5) and (10.6), which are always the same no matter what the particular machine may be, and no matter what the value of n may be.

Since the only 'missing' arrow is of form  the sentence H will be $\exists t\,\exists x\,(tQ_2x\,\&\,tS_0x)$. Now in case $n = 0$, the machine halts at time 1 (i.e. time 0'), for it will then be in state q_2, scanning a blank.

Proof that $\Delta \vdash \boldsymbol{H}$. The operative member of Δ is the sentence corresponding to the arrow $\textcircled{1} \xrightarrow{B:L} \textcircled{2}$. From that sentence we derive

$$[\mathbf{0} = x' \,\&\, \mathbf{0}Q_1\mathbf{0} \,\&\, \mathbf{0}S_0\mathbf{0}] \rightarrow [\mathbf{0}'Q_2x \,\&\, (\mathbf{0}S_0x \rightarrow \mathbf{0}'S_0x)].$$

From the description of time 0 (with $n = 0$) we derive

$$\mathbf{0}Q_1\mathbf{0} \,\&\, \mathbf{0}S_0\mathbf{0} \,\&\, (\mathbf{0} = x' \rightarrow \mathbf{0}S_0x).$$

Then we have

$$\mathbf{0} = x' \rightarrow \mathbf{0}'Q_2x \,\&\, \mathbf{0}'S_0x$$

(where the whole is understood to be governed by '$\forall x$'). By (10.5) we have

$$\exists x\, \mathbf{0} = x'.$$

These last two sentences imply

$$\exists x(\mathbf{0}'Q_2x \,\&\, \mathbf{0}'S_0x)$$

from which H follows by existential generalization.

A further exercise

10.3 A sentence S is called *finitely* satisfiable if $\mathscr{I}(S) = 1$ for some interpretation $\mathscr{I}$ whose domain is finite. Show that for each machine M and input n a sentence S can effectively be found which is finitely satisfiable iff M halts with input n. Hint: modify Δ to obtain a suitable finite set of sentences Δ' such that the conjunction of sentences in Δ' and H is finitely satisfiable iff M halts with input n. You will probably want to discard $'$ and replace it with a new two-place predicate letter ('P', say for 'predecessor of'). (10.3) might then be revised to read

$$\forall t\forall x\forall y\{[tQ_ix \,\&\, tS_jx] \rightarrow \exists z\exists w[tPz \,\&\, wPx \,\&\, zQ_mw \\ \&\, (tS_0y \rightarrow zS_0y) \,\&\, \ldots \,\&\, (tS_ry \rightarrow zS_ry)]\}.$$

(10.1) and (10.2) would be similarly revised. (10.4) would contain lots of existential quantifiers. (10.5) might read

$$\forall x\forall y\forall z(xPy \,\&\, xPz \rightarrow y = z) \,\&\, \forall x\forall y\forall z(yPx \,\&\, zPx \rightarrow y = z).$$

The middle conjunct of (10.6) might be replaced by

$$\forall x\forall y(xPy \rightarrow x < y).$$

Complete the proof.

11

First-order logic formalized: derivations and soundness

In Chapter 9 we saw that every sentence is equivalent to one in prenex normal form, and, indeed, saw how to find a prenex equivalent of any given sentence in an effective manner. We now use this fact to *formalize* first-order logic: to provide a sound, complete mechanical procedure for demonstrating the validity of valid sentences (= a mechanical positive test for validity). *Soundness* means that if the procedure classifies a sentence as valid, then the sentence really is valid. *Completeness* is defined conversely: if a sentence is valid, the procedure will so classify it. In this chapter we shall begin describing the procedure and prove a theorem (the soundness theorem) of which a consequence will be that the procedure is sound. In the next chapter we shall finish the description and prove completeness. Some important consequences of completeness are drawn in the next chapter.

A test for validity of sentences is readily modified so as to yield a test for the unsatisfiability of finite sets of sentences, for as is clear from the definitions of Chapter 9, a finite set Δ is unsatisfiable if and only if the negation of the conjunction of its members is valid. And conversely, a test for unsatisfiability is readily modified so as to yield a test for validity, for S is valid if and only if $-S$ is unsatisfiable, if and only if $\{-S\}$ is unsatisfiable. Here we shall treat the notion of unsatisfiability as basic; we will provide a positive test for unsatisfiability of finite sets of sentences which is sound (the set is unsatisfiable if the test so classifies it) and complete (the test does so classify it if it is unsatisfiable).

Note that we are now in a position to conclude that there can't be a sound, complete positive test for satisfiability (= a sound, complete negative test for unsatisfiability). For by the main result of the last chapter, a given machine fails to halt on a given input if and only if a certain sentence (the conjunction of the negation of H and the members of Δ) is satisfiable. So a positive test for satisfiability would yield a negative test for halting. But of course a mechanical *positive* test for halting does exist, *viz., imitate the operations of the machine when supplied with the input*! So if there were a mechanical positive test for satisfiability, there would be both a positive and a negative test for halting, from which it would follow that the halting problem was solvable.

The positive test for unsatisfiability of Δ will be a systematic search for a finite *refutation* of Δ, *viz*. a certain sort of finite list of sentences. Soundness of the test will come to this: if there is a finite refutation of Δ, then Δ is unsatisfiable. And completeness will come to this: if Δ is unsatisfiable, then there is a finite refutation of Δ. After we define 'refutation of Δ' (soon!) it will become apparent that if Δ is finite, or even if Δ only has the weaker property that membership in it is effectively decidable, then the property of being a finite refutation of Δ is also effectively decidable. Then a straightforward (but inefficient) way of systematically searching for a refutation of (effectively decidable) Δ would be to choose (*ad lib.*) some effective enumeration of the finite lists of sentences, and test those lists in turn to see whether they are refutations of Δ. The proof of soundness will assure us that if we do find a refutation of Δ in this way, we can be sure that Δ is unsatisfiable. And the proof of completeness will assure us that if Δ is unsatisfiable, then for some finite *n*, we shall discover that the *n*th list in our enumeration will prove to be a refutation of Δ. But if the set Δ is satisfiable, there may be no end to our search for a refutation. Zeus, testing lists of sentences faster and faster for the property of being a refutation of Δ, could get through the whole enumeration in a finite period of time, and thus establish the satisfiability of Δ by an exhaustive search which fails; but that technique is not open to us or our machines, and in Chapter 10 we saw that no other technique can work for us either – if Church's thesis is true. Of course, even where a refutation exists, it may occur so late in the chosen enumeration of all finite lists of sentences as to be inaccessible in practice: the completeness of the test guarantees that where Δ is unsatisfiable, list number *n* will be a refutation for some finite *n*, but it guarantees absolutely nothing about the size of *n*. We shall have more to say about that later.

Meanwhile, let us begin describing the test, and prove its soundness. In what follows, we don't assume that Δ is finite, or even that membership in Δ is effectively decidable. To clarify the description it will be well to have an example in mind, e.g., the inference

$$\exists x \forall y\, xLy \vdash \forall y \exists x\, xLy$$

which is valid if and only if the set consisting of the sentences

$$\exists x \forall y\, xLy, \quad -\forall y \exists x\, xLy$$

is unsatisfiable. Putting the second of these into prenex normal form, we have, as the set Δ which is to be tested for unsatisfiability, the set consisting of the sentences in lines 1 and 2 of the following annotated list.

1	$\exists x \forall y\, xLy$	Δ
2	$\exists y \forall x - xLy$	Δ
3	$\forall y\, aLy$	1
4	$\forall x - xLb$	2
5	$-aLb$	4
6	aLb	3

This list is a refutation of Δ. A refutation of Δ is a special sort of *derivation from* Δ, *viz*., one in which

> some finite set of quantifier-free lines is unsatisfiable.

(In our example, lines 5 and 6 make up an unsatisfiable set.) A derivation from Δ is defined as

> a finite or infinite list of sentences which can be annotated by writing either 'Δ' or the number of an earlier line at the right of each line, in such a way that where the annotation is 'Δ', the annotated sentence is a member of Δ, while if the annotation is *m*, the annotated sentence is obtainable from the *m*th sentence by one of the rules UI and EI which are defined below.

UI ('Universal instantiation'). The earlier line (the *premise*) is of form

$$m \quad \forall v F$$

with some annotation, and the later line (the *conclusion*) is of form

$$n \quad F_v t \quad m,$$

where *t* (the '*instantial term*') may be any term, e.g. *a* or $f(a, g(b, f(b)))$ or *b*.

EI ('Existential instantiation'). The two lines are as in the statement of UI except that we have '∃' in place of '∀' in line *m*, and the instantial term *t* in line *n* must be *a name which appears in no sentence of* Δ *and in no line earlier than n in the derivation.*

In our example, line 3 is annotated '1' and since line 1 begins with an existential quantifier, the operative rule must be EI, with $v = x$, $F = \forall y\, xLy$, and $t = a$. The instantial term appears nowhere in Δ and (therefore) nowhere earlier than line 3 of this derivation. Similarly, line 2 must be the premise of an application of EI of which line 4 is the conclusion, with the instantial term b (a name) appearing nowhere earlier than line 4 of this derivation (and thus, nowhere in Δ, since in this case, the sentences of Δ are the first two lines of the derivation). Finally, lines 5 and 6 are conclusions by UI of lines 4 and 3 respectively, with a and b respectively as instantial terms.

Here is an example of a list of sentences that is *not* a derivation from $\{\forall x \exists y\, xLy,\ \forall y \exists x - xLy\}$:

1 $\forall x \exists y\ xLy$
2 $\forall y \exists x\ -xLy$
3 $\quad \exists y\ aLy$
4 $\qquad aLb$
5 $\quad \exists x\ -xLb$
6 $\qquad -aLb$

This list is not a derivation because the sixth sentence in it, '$-aLb$' could have been inferred only by an application of EI from the fifth sentence, '$\exists x\ -xLb$'. The instantial term in this application of EI would then have been 'a'. But 'a' occurs in the earlier sentence 'aLb'. Lines 1 through 5 form a perfectly admissible derivation, however.

UI is a perfectly ordinary rule of inference, for the conclusion $F_v t$ always follows from the premise $\forall v\, F$, no matter what the instantial term t may be. Thus, if $\mathscr{I}$ is an interpretation of both the premise and the conclusion, the conclusion will be true in $\mathscr{I}$ if the premise is. But EI appears to have been stated backwards. Not only does the conclusion $F_v t$ of an application of EI not necessarily follow from the premise, the premise always follows from the conclusion! But although EI is not sound, it serves admirably as an ingredient in our sound, complete test for unsatisfiability, as long as the restriction is observed, that the instantial term t be a 'new' name – a name which appears nowhere in Δ and nowhere earlier than line n in the derivation. The thought is that when we infer (say) '$\forall x\ -xLb$' from '$\exists y \forall x\ -xLy$' we are 'giving a name' to an object y whose existence is asserted in the premise. In some interpretation $\mathscr{I}$ the premise assures us that there is someone whom no one loves. In the conclusion, we identify one such unfortunate. Since 'b' plays no role in Δ and appears nowhere earlier in the derivation (nowhere earlier than line 4 in our example), no assumptions about the bearer of the name 'b' have been formulated *except* for the assumption that he or she is totally unloved. The premise assures us that *that* assumption is true of someone, so that if the premise is true in some interpretation $\mathscr{I}$, we can be sure that somewhere in the domain of $\mathscr{I}$ there will be an object o which 'b' can be made to denote in the near variant $\mathscr{I}^b_o$ of $\mathscr{I}$; and the conclusion, which may be assigned neither truth value by $\mathscr{I}$, will be assigned the truth value 1 by $\mathscr{I}^b_o$, provided only that the premise is assigned the value 1 by $\mathscr{I}$.

In general terms, the *basic property* of EI is this:

> Suppose that all members of a set Γ of sentences which includes the premise of an application of EI are true in an interpretation $\mathscr{I}$, and suppose that the instantial term of that application is a name t which does not occur in any sentence of Γ. Then there is an object o in the domain of $\mathscr{I}$ such that in the interpretation $\mathscr{I}^t_o$, every member of Γ is true, as is the conclusion of the application of EI.

(Recall that in general, $\mathscr{I}^t_o$ is the interpretation which differs from $\mathscr{I}$ – if at all – only in that it assigns the designation o to t.) In proving that EI has this basic property we shall use the following seemingly trivial fact about interpretations.

> *Continuity.* A sentence has the same truth value in any two interpretations which differ only in what (if anything) they assign to names, sentence letters, predicate letters, or function symbols which do *not* occur in the sentence.

Proof that EI has the basic property. Since all of the sentences in Γ are true in $\mathscr{I}$, and t occurs in none of them, continuity assures us that they are all true in an interpretation $\mathscr{J}$ which differs from $\mathscr{I}$ (if at all) in assigning *no* designation to t. So the premise, $\exists vF$, is true in $\mathscr{J}$. Then by case 7 in Chapter 9, there will be an object o in the domain of $\mathscr{J}$ (= the domain of $\mathscr{I}$) such that the conclusion $F_v t$ is true in $\mathscr{J}^t_o$; and by continuity, so are all the other sentences in Γ. But $\mathscr{J}^t_o$ is just the same interpretation as $\mathscr{I}^t_o$.

We are now ready to prove the

Soundness theorem

If there is a refutation of Δ then Δ is unsatisfiable.

Here, 'Δ' is a variable for sets of sentences in prenex normal form in which no vacuous quantifiers occur. ($\exists v$ or $\forall v$ is *vacuous* if and only if the variable v has no free occurrences in what follows the quantifier, e.g.

'$\exists x$' in '$\exists x p$' or '$\forall x$' in '$\forall x \exists x Gx$'.) Since every sentence has a prenex equivalent, and vacuous quantifiers can be dropped to get sentences equivalent to the original ones, these assumptions reflect no real restriction.

The soundness theorem follows from the

Strong soundness theorem

If $\mathscr{I}$ is a model of Δ and $\mathscr{D}$ is a derivation from Δ then the set of all sentences occurring in $\mathscr{D}$ has a model $\mathscr{L}$. Moreover, $\mathscr{L}$ can be chosen so that it differs from $\mathscr{I}$ (if at all) only in what designations or functions it assigns to those names or function symbols which occur in sentences of $\mathscr{D}$ that do not belong to Δ.

It is immediate from the first sentence of the strong soundness theorem that if Δ has a refutation, then Δ is unsatisfiable. (If Δ were satisfiable, it would have a model $\mathscr{I}$, and therefore the set of all sentences occurring in the refutation would have a model $\mathscr{L}$, which is impossible as some finite subset of them is unsatisfiable.) The strong soundness theorem thus implies the soundness theorem.

Proof of the strong soundness theorem. Suppose that $\mathscr{I}$ is a model of Δ. We define $\Delta_0 = \Delta$ and in general, if $\mathscr{D}$ has an nth line, we define

$$\Delta_n = \Delta \cup \{S_1, \ldots, S_n\}$$

where the Ss are the sentences in the first n lines of $\mathscr{D}$. (Note that the Ss need not be distinct.) We consider $\mathscr{D}$ together with an arbitrary annotation which meets the requirements laid down in the definition of 'derivation from Δ':

$$\begin{array}{ccc} 1 & S_1 & A_1 \\ \vdots & \vdots & \vdots \\ n & S_n & A_n \\ \vdots & \vdots & \vdots \end{array}$$

Thus, A_1 must be 'Δ'; A_2 may be 'Δ' or '1', depending on what Δ, S_1, and S_2 are; and so on. (We speak of *an* annotation because there are derivations which can be annotated in more than one way.) Now for each n for which there is an nth line in $\mathscr{D}$ we define an interpretation $\mathscr{I}_n$ which is a model of Δ_n. Our definition will be recursive. To begin we set

$$\mathscr{I}_0 = \mathscr{I}.$$

To say that $\mathscr{I}_0$ is a model of Δ_0 is to repeat, in different words, our assumption that $\mathscr{I}$ is a model of Δ. Suppose now that we have already defined a model $\mathscr{I}_k$ of Δ_k, and that there is a $(k+1)$st line in $\mathscr{D}$. We define a model $\mathscr{I}_{k+1}$ of Δ_{k+1} in one way or another, depending on the annotation A_{k+1} and perhaps also on the particulars of the instantial term.

Case 1. $A_{k+1} =$ 'Δ'. Here we define $\mathscr{I}_{k+1} = \mathscr{I}_k$. Since $\Delta_{k+1} = \Delta_k$ in this case, we can be sure that $\mathscr{I}_{k+1}$ is a model of Δ_{k+1} for (to say the same thing in different words) we know that $\mathscr{I}_k$ is a model of Δ_k.

Case 2. A_{k+1} refers to an earlier line which begins with a universal quantifier (so that the operative rule must have been UI) and the instantial term in S_{k+1} contains only names or function symbols which occur in sentences of Δ_k. Then the model $\mathscr{I}_k$ of Δ_k is an interpretation of S_{k+1} and thus of Δ_{k+1}; and since the conclusion of an application of UI is implied by its premise (= is true in each of its interpretations in which its premise is true) then if we set $\mathscr{I}_{k+1} = \mathscr{I}_k$ we can be sure that $\mathscr{I}_{k+1}$ is a model of Δ_{k+1}.

Case 3. Like case 2 except that the instantial term in S_{k+1} contains one or more names or function symbols which occur in no sentences of Δ_k. Here we cannot rely on $\mathscr{I}_k$ to be an interpretation of S_{k+1}, although it may be. (It *will* be if and only if it happens to assign denotations and functions to the 'new' names and function symbols in the instantial term.) Now any interpretation which differs from $\mathscr{I}_k$ (if at all) only in the denotations and functions which it assigns to the 'new' names and function symbols in S_{k+1} will (by continuity) be a model of Δ_k and will (since the premise of an application of UI implies the conclusion) be a model of S_{k+1}. For definiteness, we define $\mathscr{I}_{k+1}$ as follows, where d is some element of the domain of $\mathscr{I}$ which we think of as having been chosen, once and for all instances of case 3 at the beginning of the construction. To each 'new' name, $\mathscr{I}_{k+1}$ assigns d as denotation, and to each 'new' function symbol it assigns the constant function which assumes the value d for every argument in the domain of $\mathscr{I}_k$; but $\mathscr{I}_{k+1}$ differs from $\mathscr{I}_k$ only in these assignments, if at all. (It may just happen that $\mathscr{I}_k$ itself makes those very assignments.)

Case 4. A_{k+1} refers to an earlier line which begins with an existential quantifier (so that the operative rule must have been EI). In this case the instantial term t is a name which appears in no sentence of Δ_k. As the premise of this application of EI belongs to Δ_k, it must be true in $\mathscr{I}_k$. Then by the basic property of EI, there will be at least one object o in the domain of $\mathscr{I}_k$ for which the interpretation $\mathscr{I}_k{}^t_o$ is a model of S_{k+1} and (by continuity) of Δ_k as well. We therefore define $\mathscr{I}_{k+1} = \mathscr{I}_{ko}^{t}$ for some such o.†

In all four cases, then, $\mathscr{I}_{k+1}$ is a model of Δ_{k+1}. Note that all of the $\mathscr{I}_k$s have the same domain, *viz.*, the domain of $\mathscr{I}_0$ $(= \mathscr{I})$.

Now define $\mathscr{L}$ as the interpretation which is just like $\mathscr{I}$ except that to each name or function symbol which appears in $\mathscr{D}$ but not in Δ, $\mathscr{L}$ assigns whatever $\mathscr{I}_n$ assigns it – where S_n is the earliest sentence in $\mathscr{D}$ in which the name or function symbol occurs.† Each sentence S_k occurring in $\mathscr{D}$ is then true in an interpretation $\mathscr{I}_k$ from which $\mathscr{L}$ differs only in what (if anything) it assigns to names and function symbols not in Δ_k, and so not in S_k. By continuity, therefore, each sentence in $\mathscr{D}$ is true in $\mathscr{L}$. And $\mathscr{L}$ differs from $\mathscr{I}$ (if at all) only in what it assigns to names and function symbols that appear in $\mathscr{D}$ but not in Δ.

This completes our proof of the strong soundness theorem. We have as an easy consequence that

> An inference is valid if there is a refutation of a set each of whose members is a prenex equivalent of a premise or of the denial of the conclusion of the inference.

The converse follows from the completeness theorem of Chapter 12.

A last note: To prove that it is always a mechanical matter to answer the question, 'Is $\mathscr{D}$ a refutation of Δ?' where Δ is assumed to be effectively decidable and $\mathscr{D}$ to be finite we must prove that there is a decision procedure for determining whether a set of quantifier-free sentences is satisfiable. The truth-table test alone will not do the job, since (as Exercises 11.2, 11.3, and 11.4 show) unsatisfiability of such a set may depend essentially on laws of identity. A proof that there is such a decision procedure is given at the end of the following chapter.

Exercises

Verify that each of the following inferences is valid by finding a refutation of an appropriate set of sentences.

11.1 $\forall x\, xLf(x) \vdash \forall x \exists y\, xLy$.

11.2 $\exists x \forall y (Py \leftrightarrow y = x) \vdash \exists x\, Px$.

11.3 $\exists x \forall y (Py \leftrightarrow y = x) \vdash \forall x \forall y [(Px \,\&\, Py) \rightarrow x = y]$.

11.4 $\exists x\, Px, \forall x \forall y [(Px \,\&\, Py) \rightarrow x = y] \vdash \exists x \forall y (Py \leftrightarrow y = x)$.

† Cognoscenti will recognize that we have tacitly appealed to the axiom of (dependent) choice at these points in the proof. (Cf. Exercise 11.5.)

Solutions

11.1	1	$\exists x \forall y\, xLy$	Δ
	2	$\forall y\, aLy$	1
	3	$aLf(a)$	2
	4	$\forall x\, xLf(x)$	Δ
	5	$aLf(a)$	4

$$\Delta = \{\forall x\, xLf(x), \exists x \forall y\, xLy\}.$$

The derivation above is a refutation of Δ since the set consisting of the quantifier-free sentences in lines 3 and 5 is unsatisfiable. Since the members of Δ are prenex equivalents of the premise or of the denial of the conclusion of the inference, this refutation of Δ establishes the validity of the inference.

11.2	1	$\exists x \forall y (Py \leftrightarrow y = x)$	Δ
	2	$\forall x - Px$	Δ
	3	$\forall y (Py \leftrightarrow y = a)$	1
	4	$Pa \leftrightarrow a = a$	3
	5	$-Pa$	2

$\Delta = \{S_1, S_2\}$ in the derivation above. The derivation is a refutation because the set $\{S_4, S_5\}$ is unsatisfiable (because '$a = a$' is valid).

11.3	1	$\exists x \forall y (Py \leftrightarrow y = x)$	Δ
	2	$\exists x \exists y (Px \,\&\, Py \,\&\, x \neq y)$	Δ
	3	$\forall y (Py \leftrightarrow y = a)$	1
	4	$\exists y (Pb \,\&\, Py \,\&\, b \neq y)$	2
	5	$Pb \,\&\, Pc \,\&\, b \neq c$	4
	6	$Pb \leftrightarrow b = a$	3
	7	$Pc \leftrightarrow c = a$	3

This derivation is a refutation of Δ $(= \{S_1, S_2\})$ because the last three (quantifier-free) lines form an unsatisfiable set.

11.4	1	$\exists x\, Px$	Δ
	2	$\forall x \forall y [(Px \,\&\, Py) \rightarrow x = y]$	Δ
	3	$\forall x \exists y (Py \leftrightarrow y \neq x)$	Δ
	4	Pa	1
	5	$\exists y (Py \leftrightarrow y \neq a)$	3
	6	$Pb \leftrightarrow b \neq a$	5
	7	$\forall y [(Pb \,\&\, Py) \rightarrow b = y]$	2
	8	$(Pb \,\&\, Pa) \rightarrow b = a$	7

The above is a refutation of $\{S_1, S_2, S_3\}$.

To see that $\{S_4, S_6, S_8\}$ is unsatisfiable, note that by sentential logic it implies '$-Pb$' and '$b = a$', whence (substituting equals for equals) '$-Pa$', which contradicts S_4.

Exercise 11.5

(For those worried about the use of the axiom of dependent choice in the soundness proof.)

The axiom of dependent choice asserts that if X is a non-empty set, and for any x in X there is a y in X such that x bears relation R to y, then there is a function f such that for any natural number n, $f(n)$ is in X and $f(n)$ bears R to $f(n+1)$. This assertion is not to be confused with the weaker assertion whose antecedent is the same, but whose consequent is 'for any natural number n, there is a function f such that for any natural number $i < n$, $f(i)$ is in X and $f(i)$ bears R to $f(i+1)$. Unlike the axiom of dependent choice, this second assertion can be proved (in set theory) without any extra assumptions. Since the strong soundness theorem is equivalent in set theory to the axiom of dependent choice, uses of that axiom can only be disguised, not essentially avoided, in any proof of the theorem.

Show that the soundness theorem (as opposed to the strong soundness theorem) can be proved without appeal to the axiom of dependent choice. *Hint*: use the fact that if there is a refutation of Δ, there is one which contains only finitely many sentences. A proof of the result of modifying the strong soundness theorem by inserting the words 'finitely long' before the word 'derivation' in its statement may be given in which appeal need be made only to the weaker statement mentioned above instead of the axiom of dependent choice. In the original proof we needed the axiom of dependent choice to guarantee us of the existence of $\mathscr{L}$ if $\mathscr{D}$ was infinite; if $\mathscr{D}$ is assumed to be finite, we need only the truth of the weaker statement.

12
Completeness of the formalization; compactness

We now prove the

Completeness theorem

If Δ† is unsatisfiable, there is a refutation of Δ.

An analysis of the completeness proof will reveal two important facts. First, the

Compactness theorem

If Δ is unsatisfiable, some finite subset of Δ must be unsatisfiable.

Second, the

Skolem–Löwenheim theorem

If Δ has a model, it has a model with an enumerable domain.

The ramifications of this second fact are sufficiently striking to warrant extensive treatment: see Chapter 13.

We prove the completeness theorem by showing how to generate special sorts of derivation – *canonical* derivations – which have the characteristic that

If Δ is unsatisfiable, any canonical derivation from Δ will be a refutation of Δ.

Definition

$\mathscr{D}$ is a *canonical derivation* from Δ if and only if $\mathscr{D}$ is a derivation from Δ which has these five characteristics:

(1) Every sentence in Δ occurs in $\mathscr{D}$.

(2) If a sentence $\exists v\, F$ occurs in $\mathscr{D}$, then for some term t, the sentence $F_v t$ occurs in $\mathscr{D}$.

(3) If a sentence $\forall v\, F$ occurs in $\mathscr{D}$, then for some term t, the sentence $F_v t$ occurs in $\mathscr{D}$.

(4) If a sentence $\forall v\, F$ occurs in $\mathscr{D}$, then for every term t that can be

† Here and henceforth, sets of sentences are always assumed to be enumerable.

To see that $\{S_4, S_6, S_8\}$ is unsatisfiable, note that by sentential logic it implies '$-Pb$' and '$b = a$', whence (substituting equals for equals) '$-Pa$', which contradicts S_4.

Exercise 11.5

(For those worried about the use of the axiom of dependent choice in the soundness proof.)

The axiom of dependent choice asserts that if X is a non-empty set, and for any x in X there is a y in X such that x bears relation R to y, then there is a function f such that for any natural number n, $f(n)$ is in X and $f(n)$ bears R to $f(n+1)$. This assertion is not to be confused with the weaker assertion whose antecedent is the same, but whose consequent is 'for any natural number n, there is a function f such that for any natural number $i < n$, $f(i)$ is in X and $f(i)$ bears R to $f(i+1)$. Unlike the axiom of dependent choice, this second assertion can be proved (in set theory) without any extra assumptions. Since the strong soundness theorem is equivalent in set theory to the axiom of dependent choice, uses of that axiom can only be disguised, not essentially avoided, in any proof of the theorem.

Show that the soundness theorem (as opposed to the strong soundness theorem) can be proved without appeal to the axiom of dependent choice. *Hint*: use the fact that if there is a refutation of Δ, there is one which contains only finitely many sentences. A proof of the result of modifying the strong soundness theorem by inserting the words 'finitely long' before the word 'derivation' in its statement may be given in which appeal need be made only to the weaker statement mentioned above instead of the axiom of dependent choice. In the original proof we needed the axiom of dependent choice to guarantee us of the existence of $\mathscr{L}$ if $\mathscr{D}$ was infinite; if $\mathscr{D}$ is assumed to be finite, we need only the truth of the weaker statement.

12
Completeness of the formalization; compactness

We now prove the

Completeness theorem

If Δ† is unsatisfiable, there is a refutation of Δ.

An analysis of the completeness proof will reveal two important facts. First, the

Compactness theorem

If Δ is unsatisfiable, some finite subset of Δ must be unsatisfiable.

Second, the

Skolem–Löwenheim theorem

If Δ has a model, it has a model with an enumerable domain.

The ramifications of this second fact are sufficiently striking to warrant extensive treatment: see Chapter 13.

We prove the completeness theorem by showing how to generate special sorts of derivation – *canonical* derivations – which have the characteristic that

If Δ is unsatisfiable, any canonical derivation from Δ will be a refutation of Δ.

Definition

$\mathscr{D}$ is a *canonical derivation* from Δ if and only if $\mathscr{D}$ is a derivation from Δ which has these five characteristics:

(1) Every sentence in Δ occurs in $\mathscr{D}$.

(2) If a sentence $\exists v\, F$ occurs in $\mathscr{D}$, then for some term t, the sentence $F_v t$ occurs in $\mathscr{D}$.

(3) If a sentence $\forall v\, F$ occurs in $\mathscr{D}$, then for some term t, the sentence $F_v t$ occurs in $\mathscr{D}$.

(4) If a sentence $\forall v\, F$ occurs in $\mathscr{D}$, then for every term t that can be

† Here and henceforth, sets of sentences are always assumed to be enumerable.

formed from names and function symbols appearing in 𝒟, the sentence $F_v t$ occurs in 𝒟.

(5) All function symbols appearing in 𝒟 appear in Δ.

To generate a canonical derivation from Δ, follow the instructions given in the proof of the following

Lemma I

For any set Δ, there is a canonical derivation 𝒟 from Δ.

Proof. Let F_1, F_2, ... be an enumeration of all the sentences of Δ, if any. (Δ may be empty. If it is not, the enumeration is to be a gapless list – finite or infinite.) If Δ is empty, 𝒟 will be the vacuous list. Otherwise, we form 𝒟 in a series of stages. At each stage, a finite number of lines will be added to 𝒟. There will be infinitely many stages, even if Δ is finite. Each stage will have three parts.

Stage 1a. Enter

$$1 \quad F_1 \quad \Delta$$

as the first line of 𝒟.

Stage 1b. Extend 𝒟 by adding to it as many lines as can possibly be inferred by EI under certain restrictions (stated below).

Stage 1c. Extend 𝒟 further by adding to it as many lines as can possibly be inferred by UI (under certain restrictions).

Stage 2a. Enter

$$n \quad F_2 \quad \Delta$$

as the next line (where n is the appropriate number) if there is a second entry in the enumeration of Δ (under restrictions).

Stage 2b. Extend 𝒟 by adding to it as many lines as can possibly be inferred by EI (under restrictions).

Stage 2c. Extend 𝒟 further by adding to it as many lines as can possibly be inferred by UI (under restrictions).

And so on.

The restrictions are these:

We never enter the same sentence in two different lines of 𝒟.

No sentence may be the premise of more than one application of EI.

Whenever, during part *c* of stage *N*, we apply UI, the instantial term is always one which *contains fewer than N occurrences of function symbols* and which can be formed from names and function symbols already occurring in sentences of 𝒟 (*except* in the one case where F_N begins with a universal quantifier and no names appear in 𝒟 as yet. In that case we apply UI to F_N using as the instantial term any name we please.)

Of course, in part *b* of stage *N*, new sentences may be added to 𝒟 which can themselves be the premises of applications of EI; and any such sentences are supposed to be used *in part b of stage N* as the premises of applications of EI. Similarly for part *c* of stage *N*, and UI. During each part of each stage, only finitely many sentences are added to the derivation: we always eventually run out of terms to substitute or sentences to use as premises or both. (That is precisely the point of the third restriction, e.g. in part *c* of stage 3 we could use only the first three members of the series '*a*', '*f*(*a*)', '*f*(*f*(*a*))', '*f*(*f*(*f*(*a*)))', ... as instantial terms.)

Example

Δ = {F_1, F_2}, where F_1 = '∀*x xLfx*' and F_2 = '∃*x*∀*y xŁy*', as in the solution to Exercise 11.1. Here is the part of a canonical derivation from Δ which is yielded by the first two stages of our procedure.

1	∀*x xLfx*	Δ	(1*a*)
2	*aLfa*	1	(1*c*)
3	∃*x*∀*y xŁy*	Δ	(2*a*)
4	∀*y bŁy*	3	(2*b*)
5	*bLfb*	1	(2*c*)
6	*faLffa*	1	(2*c*)
7	*fbLffb*	1	(2*c*)
8	*bŁa*	4	(2*c*)
9	*bŁb*	4	(2*c*)
10	*bŁfa*	4	(2*c*)
11	*bŁfb*	4	(2*c*)

Note that at stage 2*c*, the allowable instantial terms are those with fewer than two occurrences of '*f*': '*a*', '*b*', '*fa*' and '*fb*'. All of these were used in lines 8–11, and all but '*a*' were used in lines 5–7. (Use of '*a*' there would have produced a replica of line 2, in violation of the first restriction.) At stage 3, only part *c* will be applicable: UI will be applied to lines 1 and 4, using as instantial terms in each case whichever of '*a*', '*b*' '*fa*', '*fb*', '*ffa*', and '*ffb*' do not produce violations of the first restriction. Although these first 11 lines make up a refutation of Δ (since lines 5 and 11 form an unsatisfiable set), they do not constitute a canonical derivation. Any canonical derivation from Δ in this case will be unending, for at each stage, application of UI to line 1 will generate new terms which provide fuel for the next stage.

The next concept we shall need has to do with interpretations $\mathscr{I}$ and sets Γ of *quantifier-free* sentences:

> $\mathscr{I}$ *matches* Γ if and only if
> $\mathscr{I}$ is a model of Γ, and
> if any terms at all occur in Γ† then each object in the domain of $\mathscr{I}$ is the denotation of some such term.

The case in which no terms occur in Γ is simply the case in which Γ is a set of sentences built out of sentence letters '*p*' etc., so that the domain of $\mathscr{I}$ is irrelevant to the truth-values which $\mathscr{I}$ assigns to members of Γ. Of course, if $\mathscr{I}$ is any model of Γ, every term occurring in a sentence of Γ must denote something or other in the domain of $\mathscr{I}$. But where $\mathscr{I}$ matches Γ, the converse holds as well: every object in the domain of $\mathscr{I}$ is denoted by one or another term which appears in a sentence of Γ.

Where Γ is the set of all quantifier-free sentences in a canonical derivation from Δ, we can rely on any interpretation which matches Γ to be a model of Δ:

Lemma II

Suppose that $\mathscr{D}$ is a canonical derivation from Δ, that Γ is the set of all quantifier-free sentences in $\mathscr{D}$, and that $\mathscr{I}$ matches Γ. Then $\mathscr{I}$ is a model of the set of *all* sentences in $\mathscr{D}$, and hence is a model of Δ.

Proof. Since every non-logical symbol, i.e. name, function symbol, sentence letter or predicate letter that appears in $\mathscr{D}$ appears in Γ, $\mathscr{I}$ will assign a truth-value to each sentence in $\mathscr{D}$. To prove the lemma it suffices to prove that in no case is the truth-value 0 (falsity). To prove this by *reductio ad absurdum*, suppose $\mathscr{I}$ assigns the value 0 to one or more sentences in $\mathscr{D}$. Then among the lengths of all such sentences, there must be a minimum, say, *m*. (The length of a sentence is the number of symbols in it, *counting terms as single symbols*.) Let *M* be some sentence of length *m* in $\mathscr{D}$, to which $\mathscr{I}$ assigns the value 0. Since *M* cannot be quantifier-free, *M* must begin with a quantifier. *If the quantifier is existential*, then by clause (2) of the definition of 'canonical derivation', some instance of *M*, shorter (of length $m-2$) and hence true in $\mathscr{I}$, occurs in $\mathscr{D}$; but then *M* would be true in $\mathscr{I}$, as any instance implies it. *If the quantifier is universal*, then by clause (4) and the fact that every object in the domain of $\mathscr{I}$ is denoted by some term appearing in Γ, some instance of *M* of which the instantial term appears in Γ must occur in $\mathscr{D}$ and be false in $\mathscr{I}$ – which is impossible, since instances of *M* are shorter than *M* and hence are true in $\mathscr{I}$ if they occur in $\mathscr{D}$.

† A term is understood to occur in Γ even if it occurs only as a part of some other term that occurs in Γ.

Let us pause for a moment to see how far we have come. We are trying to prove that if Δ is unsatisfiable, then there is a canonical derivation from Δ in which some finite number of quantifier-free sentences form an unsatisfiable set. So let us suppose that Δ is unsatisfiable. By Lemma I, there is a canonical derivation $\mathscr{D}$ from Δ. Since Δ is unsatisfiable and is a subset of the set of all sentences appearing in $\mathscr{D}$, there is no model of the set of all such sentences. By Lemma II, then, there is no model of the set Γ of quantifier-free sentences in $\mathscr{D}$ that matches Γ. If we could prove a proposition to the effect that

> if every finite subset of Γ is satisfiable then some interpretation matches Γ

then we should have proved the completeness theorem, for we should know that some finite set of the quantifier-free sentences in $\mathscr{D}$ is unsatisfiable.

In order to prove this proposition we shall introduce the concept of an *O.K.* set of sentences, and shall briefly discuss the concepts of an *equivalence relation* and an *equivalence class.*

Definition

A set θ of sentences is *O.K.* if and only if every finite subset of θ is satisfiable.

(Then the antecedent of the proposition which we seek to prove is that Γ is O.K.) An important fact about O.K.-ness is that

> If θ is O.K. and *S* is any sentence, then either $\theta \cup \{S\}$ is O.K. or $\theta \cup \{-S\}$ is.

Proof. Note that if each of the sets

$$\{A_1, \ldots, A_m, S\}, \quad \{B_1, \ldots, B_n, -S\}$$

is unsatisfiable, so is the set

$$\{A_1, \ldots, A_m, B_1, \ldots, B_n\}.$$

(For if this last set is satisfied by some interpretation, it is satisfied by an interpretation in which one of the sentences S, $-S$, is true, and hence in which all members of one of the former two sets is true.) So if $\theta \cup \{S\}$ is not O.K., some subset $\{A_1, \ldots, A_m, S\}$ is unsatisfiable where, we may assume, all of the As belong to θ. Similarly if $\theta \cup \{-S\}$ is not O.K., one of *its* subsets $\{B_1, \ldots, B_n, -S\}$ is unsatisfiable, where the Bs all belong to θ. Then if both fail to be O.K., a subset $\{A_1, \ldots, A_m, B_1, \ldots, B_n\}$ of θ is unsatisfiable, and thus θ is not O.K.

Equivalence relations

Suppose that X is a nonempty set and that R is a relation on X, i.e. suppose that whenever we have xRy, then x and y both belong to X.

Definition. R is an equivalence relation on X if and only if,
R is *reflexive* on X (xRx whenever x is in X),
R is *transitive* (xRz whenever xRy and yRz), and
R is *symmetrical* (xRy whenever yRx).

If R is an *equivalence relation* on X, and x is in X, then the set of those members of X to which x bears R is called *the equivalence class of x under the relation R*. A customary designation for the equivalence class of x under R is '$[x]_R$'. Usually it is clear from the context what relation R is in question, and then the subscript 'R' is generally omitted. Thus we define

$$z \in [x] \text{ if and only if } xRz.$$

We shall need the following facts about equivalence classes and equivalence relations:

If R is an equivalence relation on X and x and y are in X, then
(1) x is in $[x]$, and
(2) xRy if and only if $[x] = [y]$.

Proof of (1). R is reflexive.

Proof of (2). For the 'only if' part, suppose that xRy. If z is in $[x]$ then xRz and thus, by symmetry, zRx, so that by transitivity, zRy, and by symmetry again, yRz, so that z is in $[y]$: and if z is in $[y]$ then by completely parallel reasoning we have z in $[x]$. Therefore $[x] = [y]$. For the 'if' part, suppose that $[x] = [y]$. Since by (1), y is in $[y]$, we have y in $[x]$ and hence xRy. This completes the proof of (2).

Now as the sentences, and hence the quantifier-free sentences, in any derivation form an enumerable set, we shall have established the completeness theorem when we have proved the following proposition:

Lemma III

If Γ is an enumerable, O.K. set of quantifier-free sentences, then there is an interpretation $\mathscr{I}$ which matches Γ.

Since the proof is long, let us first outline it. We shall define a sequence of sentences $A_1, A_2, \ldots$; then, a sequence $\Gamma_1, \Gamma_2, \ldots$ of O.K. sets; and then a sequence $B_1, B_2, \ldots$, of sentences. The sequence $B_1, B_2, \ldots$ will be used to define an equivalence relation on the set of terms appearing in Γ, and this equivalence relation is then used to define $\mathscr{I}$. The Bs are then shown to be all true in $\mathscr{I}$. By the time this has been shown, it will have become clear that if $\mathscr{I}$ is a model of Γ, then $\mathscr{I}$ matches Γ. Finally, all members of Γ are shown to be true in $\mathscr{I}$.

Proof of lemma III. Suppose that Γ is an enumerable, O.K. set of quantifier-free sentences. We may assume that Γ is nonempty, for otherwise every interpretation $\mathscr{I}$ matches Γ.

Let $A_1, A_2, \ldots$ be an enumeration of all atomic sentences which are either sentence letters appearing in sentences of Γ or sentences which can be formed by filling the blanks of the equals-sign and predicate letters appearing in Γ with terms that occur in sentences in Γ. (Include sentences formed by filling the blanks of the equals-sign even when that sign does not appear in Γ.)

We now define the sequence $\Gamma_1, \Gamma_2, \ldots$ and verify that all members are O.K. Let $\Gamma_1 = \Gamma$, which is O.K. by hypothesis. Now suppose that Γ_n has been defined so as to be O.K. Then as we have noted, at least one of the sets $\Gamma_n \cup \{A_n\}$, $\Gamma_n \cup \{-A_n\}$ is O.K. We define Γ_{n+1} to be the O.K. one of the two if exactly one is O.K., and we define $\Gamma_{n+1} = \Gamma_n \cup \{A_n\}$ if both are O.K. Then Γ_{n+1} is O.K. if Γ_n is, and all members of the sequence are O.K.

It is clear from the definition of the Γ's that if $i \leqslant j$ then $\Gamma_i \subseteq \Gamma_j$, and also that $\Gamma \subseteq \Gamma_i$ for all i. Since all the Γs are O.K., it never happens that both A_i and $-A_i$ are in some Γ_j. But since at least one of A_i, $-A_i$ is in

Γ_{i+1}, *exactly one* of them is in Γ_{i+1}, and hence exactly one of them is in Γ_m, for all $m \geqslant i+1$. *We define B_i to be whichever of $A_i, -A_i$ is in Γ_{i+1}.*

We now examine the sequence of Bs rather carefully. If r and s are terms (*viz*., terms which occur in Γ – henceforth we shall usually omit this qualification), exactly one of the two sentences $r = s$, $r \neq s$ is in the sequence of Bs. (*Proof.* If both $r = s$ and $r \neq s$ were in the sequence of Bs, then for some i and j, $r = s$ would be B_i and $r \neq s$ would be B_j, and so both would be in Γ_{m+1}, where m is the larger of i and j. But then Γ_{m+1} would not be O.K., as it would have $\{r = s, r \neq s\}$ as an unsatisfiable subset.) We now define

$r \sim s$ *if and only if the sentence $r = s$ is one of the* Bs.

Note that for any term r we have $r \sim r$. (*Proof.* If B_i were $r \neq r$ then Γ_{i+1} would not be O.K., for it would have $\{r \neq r\}$ as an unsatisfiable finite subset. Then $r \neq r$ is not one of the Bs, and hence $r = r$ must be one, and so $r \sim r$.) Then the relation $\sim$ is reflexive on the set of terms. It is also transitive. (*Proof.* For *reductio ad absurdum*, suppose $r \sim s$ and $s \sim t$ but $r \nsim t$, i.e. suppose that $r = s$ is B_i, $s = t$ is B_j, but $r \neq t$ is B_k, for some i, j, k. Then all of $r = s$, $s = t$, $r \neq t$ would be in Γ_{m+1}, where m is the largest of i, j, k, and Γ_{m+1} would then not be O.K. Thus, if B_i is $r = s$ and B_j is $s = t$ then for no k is B_k $r \neq t$, and hence for some k, B_k is $r = t$, and so $r \sim t$.) Finally, the relation $\sim$ is symmetrical. (*Proof.* If $r \sim s$ then for some i, B_i is $r = s$. Then for no j is B_j $s \neq r$ and thus for some j, B_j is $s = r$, so that $s \sim r$.) We have now established that $\sim$ is an equivalence relation on the set of terms. Every term t thus belongs to a unique equivalence class $[t]$ under the relation $\sim$: a class of which each member bears $\sim$ to every other member, and which contains every term which bears $\sim$ to any one of its members.

We can now begin to define the interpretation $\mathscr{I}$ which matches Γ:

(A) If no names and predicate letters (and hence no terms) appear in Γ, the domain of $\mathscr{I}$ may be any nonempty set at all; otherwise, the domain is to be the set of all equivalence classes $[t]$.

We wish to define $\mathscr{I}$ so that $\mathscr{I}$ assigns to each term t its own equivalence class as denotation, and so that $\mathscr{I}$ makes each of the Bs true: we want to have $\mathscr{I}(t) = [t]$ and $\mathscr{I}(B_i) = 1$ for each t and i. There is essentially only one way to finish the definition of $\mathscr{I}$ so as to satisfy this wish:

(B) $\mathscr{I}$ assigns to each *name* t the equivalence class $[t]$ as its designation.

(C) $\mathscr{I}$ assigns to each n-place function symbol f the function $\not{f}$ which is determined by the condition that for all $[t_1], \ldots, [t_n]$ in the domain of $\mathscr{I}$, $\not{f}([t_1], \ldots, [t_n]) = [f(s_1, \ldots, s_n)]$ if there are terms $s_1, \ldots, s_n$ in $[t_1], \ldots, [t_n]$ respectively such that $f(s_1, \ldots, s_n)$ is a term appearing in Γ; and otherwise we set $\not{f}([t_1], \ldots, [t_n]) = [t]$ where t is any term we please appearing in Γ.

(D) $\mathscr{I}$ specifies that a sentence letter is to be true if and only if that letter is one of the Bs.

(E) $\mathscr{I}$ specifies that an n-place predicate letter R is to be true of $[t_1], \ldots, [t_n]$ (in that order) if and only if $Rt_1, \ldots, t_n$ is one of the Bs.

Now it may appear that there is something improper about clauses (C) and (E) of the definition of $\mathscr{I}$. Thus, in (C) we have specified that if there are terms $s_1, \ldots, s_n$ in $[t_1], \ldots, [t_n]$ such that $f(s_1, \ldots, s_n)$ appears in Γ, then $\not{f}([t_1], \ldots, [t_n]) = [f(s_1, \ldots, s_n)]$. But it is entirely possible that there are also terms $r_1, \ldots, r_n$ in $[t_1], \ldots, [t_n]$ such that $f(r_1, \ldots, r_n)$ appears in $\mathscr{I}$ (without, say, r_1 being identical with s_1), and then $\not{f}([t_1], \ldots, [t_n])$ must also be $[f(r_1, \ldots, r_n)]$. Then unless we have some guarantee that in this case $[f(s_1, \ldots, s_n)] = [f(r_1, \ldots, r_n)]$, we cannot claim to have determined a *unique* value of the function $\not{f}$ for the arguments $[t_1], \ldots, [t_n]$, and hence we cannot claim to have defined a *function* $\not{f}$ at all.

We have such a guarantee. Because the set

$$\{r_1 = s_1, \ldots, r_n = s_n, f(r_1, \ldots, r_n) \neq f(s_1, \ldots, s_n)\}$$

is unsatisfiable, we can conclude that if r_1 and s_1 are both in $[t_1], \ldots,$ and r_n and s_n are both in $[t_n]$, then $r_1 \sim s_1, \ldots,$ and $r_n \sim s_n$, and hence

$$f(r_1, \ldots, r_n) \sim f(s_1, \ldots, s_n)$$

and hence $[f(r_1, \ldots, r_n)] = [f(s_1, \ldots, s_n)]$.

A similar question arises in connection with clause (E): What guarantee have we that we have determined a unique truth-value for the predicate letter R with respect to $[t_1], \ldots, [t_n]$? That we have such a guarantee is shown by the observation that since the sets

$$\{s_1 = t_1, \ldots, s_n = t_n, Rs_1, \ldots s_n, -Rt_1, \ldots, t_n\}$$

and $$\{s_1 = t_1, \ldots, s_n = t_n, -Rs_1, \ldots, s_n, Rt_1, \ldots, t_n\}$$

are both unsatisfiable, if all of the conditions $s_1 \sim t_1, \ldots, s_n \sim t_n$ hold then $Rs_1, \ldots, s_n$ is one of the Bs if and only if $Rt_1, \ldots, t_n$ is.

It follows inductively from clauses (B) and (C) that each term t appearing in Γ denotes $[t]$. For by (B), each name t denotes $[t]$; and if t appears in Γ with $t = f(t_1, \ldots, t_n)$, and if $\not{f}$ is as in (C) and t_1 denotes $[t_1], \ldots$ and t_n denotes $[t_n]$, then t denotes $\not{f}([t_1], \ldots, [t_n]) = [f(t_1, \ldots, t_n)] = [t]$.

We now want to see that each of the Bs is true in $\mathscr{I}$. For each i, either $B_i = A_i$ or $B_i \neq A_i$. If $B_i \neq A_i$ then B_i is true in $\mathscr{I}$ if and only if A_i is not true in $\mathscr{I}$. We thus want to see that A_i is true in $\mathscr{I}$ if and only if $B_i = A_i$. Now A_i can be either a sentence letter, a sentence $Rt_1, \ldots, t_n$, or a sentence $s = t$. If A_i is a sentence letter then by clause (D), A_i is true in $\mathscr{I}$ if and only if $A_i = B_i$. If $A_i = Rt_1, \ldots, t_n$ then by clause (E) and the fact that each t denotes $[t]$, A_i is true if and only if $A_i = B_i$. Finally, if A_i is $s = t$ then A_i is true in $\mathscr{I}$ if and only if the denotation of s is the same as the denotation of t, i.e. $[s] = [t]$, i.e. $s \sim t$, i.e. $s = t$ is one of the Bs, i.e. $s = t$ is B_i, i.e. $A_i = B_i$.

In order to see that $\mathscr{I}$ matches Γ we have to see that in case the members of Γ are not simply built out of sentence letters and connectives, every object in the domain of $\mathscr{I}$ is the denotation (according to $\mathscr{I}$) of some term appearing in Γ, and also that $\mathscr{I}$ is a model of Γ. But where the members of Γ are not simply built out of sentence letters and connectives, the objects in the domain of $\mathscr{I}$ are just the equivalence classes of terms appearing in Γ, and we have already seen that each of these is denoted by any one of its members, which are terms that appear in Γ. So we need only show that $\mathscr{I}$ is a model of Γ.

Suppose then that S is a member of Γ. For any i, B_i is true in $\mathscr{I}$. Therefore, if B_i is true in an interpretation $\mathscr{J}$, A_i will have the same truth-value in $\mathscr{J}$ that it has in $\mathscr{I}$. S is a truth-functional compound of some finite number of the As, and therefore there is a positive integer k such that all of the As of which S is a compound are members of $\{A_1, \ldots, A_k\}$. Therefore in any interpretation $\mathscr{J}$ in which all of $B_1, \ldots, B_k$ are true, each of $A_1, \ldots, A_k$ has the same truth-value that it has in $\mathscr{I}$, and hence S has the same truth-value that it has in $\mathscr{I}$. All of $B_1, \ldots, B_k$ are in Γ_{k+1}, as is S, which is in Γ_1. Then $\{B_1, \ldots, B_k, S\}$ is a subset of Γ_{k+1}. Since Γ_{k+1} is O.K. and this subset is finite, it is satisfiable, and hence true in some interpretation $\mathscr{J}$. As all of $B_1, \ldots, B_k$ and S are true in $\mathscr{J}$, S is true in $\mathscr{I}$. Thus $\mathscr{I}$ is a model of Γ, and the completeness theorem is proved.

Let us now ponder three of the more significant consequences of the completeness theorem. First:

The compactness theorem for first-order logic

A set θ of sentences is unsatisfiable if and only if some finite subset θ_0 of θ is unsatisfiable.

Proof. The 'if' part is trivial: if *some* subset (finite or not) of θ is unsatisfiable, so is θ. For the 'only if' part, suppose θ unsatisfiable. We may assume that each member of θ is in prenex normal form. By the completeness theorem, there is a derivation $\mathscr{D}$ from θ such that some finite set $\{A_1, \ldots, A_m\}$ of quantifier-free sentences that occur in $\mathscr{D}$ is unsatisfiable. We may suppose that these sentences first occur in $\mathscr{D}$ in the order $A_1, \ldots, A_m$. Consider the derivation $\mathscr{D}_0$ that is obtained from $\mathscr{D}$ by deleting all sentences that occur in $\mathscr{D}$ after A_m. $\mathscr{D}_0$ contains only finitely many sentences, and therefore there are only finitely many members $F_1, \ldots, F_n$ of θ that occur in $\mathscr{D}_0$. Let $\theta_0 = \{F_1, \ldots, F_n\}$. Since all of $A_1, \ldots, A_m$ occur in $\mathscr{D}_0$, $\mathscr{D}_0$ is a derivation from θ_0 such that some set of quantifier-free sentences that occur in $\mathscr{D}_0$ is unsatisfiable. Therefore, by the soundness theorem, θ_0 is unsatisfiable, and is thus a finite unsatisfiable subset of θ. Upshot:

> A sentence is implied by a set of sentences if and only if it is implied by some finite subset of it.

Proof. If Γ implies S then $\Gamma \cup \{-S\}$ is unsatisfiable, whence, by the compactness theorem, some finite subset θ_0 of $\Gamma \cup \{-S\}$ is unsatisfiable. If $\Gamma_0 = \theta_0 - \{-S\}$,† Γ_0 is a finite subset of Γ and implies S, for $\Gamma_0 \cup \{-S\}$ is unsatisfiable. Second:

The Skolem–Löwenheim theorem

If a set Δ of sentences has a model, then it has a model with an enumerable domain.

Proof. Suppose that Δ has a model, i.e. is satisfiable. By Lemma I there is a canonical derivation $\mathscr{D}$ from Δ. By the soundness theorem, $\mathscr{D}$ is not a refutation of Δ and so no finite subset of the set Γ of quantifier-free sentences in $\mathscr{D}$ is unsatisfiable. So Γ is an enumerable O.K. set of sentences. Thus by Lemma III there is an interpretation $\mathscr{I}$ which matches Γ, and by Lemma II $\mathscr{I}$ is a model of Δ.

Now if Δ is a set of sentences in none of which are there any predicate letters, then all of the non-logical symbols appearing in Δ are sentence letters, and any interpretation $\mathscr{J}$, whose domain is some arbitrarily chosen enumerable set, and which assigns to the sentence letters appearing in Δ whatever truth-values $\mathscr{I}$ assigns to them will be a model of Δ with an enumerable domain.

† $A - B$ is the set of things in A that are not in B.

But if there is at least one predicate letter appearing in Δ, then there is at least one term appearing in Γ, and since $\mathscr{I}$ matches Γ, everything in the domain of $\mathscr{I}$ is the denotation of some term appearing in Γ. But as Γ is enumerable, there are at most enumerably many such terms, and hence the domain of $\mathscr{I}$ will be enumerable, and thus in this case $\mathscr{I}$ itself will be a model of Δ with an enumerable domain. Third:

There is an effective positive test for unsatisfiability (and hence, for validity)

We are now in a position to see that there is an effective positive test for unsatisfiability: an effective procedure which, when applied to an arbitrary sentence S of some first-order language, terminates with a 'yes' if and only if S is unsatisfiable. An effective positive test for unsatisfiability is not the same thing as a decision procedure for unsatisfiability, which we may take to be an effective method which terminates with a 'yes' if the sentence to which it is applied is unsatisfiable, and terminates with a 'no' if the sentence is satisfiable. (Not terminating with a 'yes' is not the same thing as terminating with a 'no', for one way of not terminating with a 'yes' is not terminating at all!) Our description of the procedure will be brief and intuitive, but it ought to be quite clear from it how one might go about actually writing down the table of a (quite large) Turing machine M^* which, when given an arbitrary sentence of some first-order language as input, halted after some finite number of steps with the words 'yes, unsatisfiable' ('yes, valid') appearing on its tape if and only if the sentence was unsatisfiable (valid).

The proof of Lemma I showed us that for any prenex sentence S there is a canonical derivation from $\{S\}$. In general, there are infinitely many canonical derivations from any set Δ. But if we make certain further decisions not already indicated in the proof of Lemma I (about which sentences and which terms are to be used in applications of UI and EI if there is a choice), we can associate with each prenex S a *unique* canonical derivation $\mathscr{D}_S$ from $\{S\}$, and can give effective instructions for writing down arbitrarily long finite initial segments of $\mathscr{D}_S$. That is to say, given any prenex sentence S and any number n, we can (effectively) find effective instructions for writing down the first n sentences of $\mathscr{D}_S$ – or the whole of $\mathscr{D}_S$, if $\mathscr{D}_S$ contains fewer than n sentences.

In detail, our effective positive test can be described as follows. *First*: find a prenex equivalent P of S. (Effective instructions can be given for finding a prenex equivalent of any given sentence.) *Second*: list longer and longer initial portions of $\mathscr{D}_P$, pausing after each sentence is written down to decide whether the set of all quantifier-free sentences so far listed is satisfiable or not (an effective procedure for doing this is given below). Stop altogether and write 'yes' if the set is unsatisfiable, and otherwise continue, writing down the next sentence in $\mathscr{D}_P$. The soundness and completeness theorems imply that this procedure eventually terminates with a 'yes' if and only if $\{P\}$ is unsatisfiable, and hence if and only if $\{S\}$ is unsatisfiable.

Our procedure will thus always give the right answer when that answer is 'yes, unsatisfiable'. But where the right answer is 'No, satisfiable', our procedure will not in general show that to be the right answer: we may continue to list sentences *ad infinitum* without ever being in a position to effectively calculate that in fact we will continue to list sentences *ad infinitum*. Moreover, this is not a defect of the particular procedure we have described, for as we saw in Chapter 10, there can be no effective negative test for unsatisfiability. But our *positive* test for unsatisfiability is premised on the existence of both positive and negative tests for unsatisfiability of finite sets of *quantifier-free* sentences. We conclude by proving that such tests exist.

How to decide whether or not a finite set of quantifier-free sentences is satisfiable. Lemma III guarantees us that there is a procedure for deciding whether a finite set Γ of quantifier-free sentences is satisfiable or not, for either (1) all members of Γ are built out of sentence letters with the aid of connectives and parentheses, in which case we can use truth tables or some other well known method to determine whether Γ is satisfiable, or (2) there is a positive integer n which is the number of distinct terms occurring in sentences of Γ. Now if Γ is satisfiable, Γ is O.K., whence by Lemma III there is a model of Γ whose domain contains no more than n members. Therefore, Γ is satisfiable if and only if it has a model whose domain contains no more than n members.

It follows that Γ is satisfiable if and only if satisfied by some interpretation $\mathscr{I}$ of which the domain is $\{1, \ldots, m\}$ (where m is some number between 1 and n, inclusive) and which specifies nothing about symbols which do not occur in Γ. There are only finitely many such interpretations and, supplied with Γ, we can effectively give instructions for writing down an explicit description – sometimes called a *diagram* – of each of them. In a diagram of $\mathscr{I}$ it is explicitly stated what number each name in Γ denotes, what the truth-value of each sentence letter in Γ is, what the value is for the function $\mathscr{I}$ assigns to each function symbol in Γ, for each sequence of arguments in the domain, and which truth-value $\mathscr{I}$ assigns to each predicate letter, for each sequence of arguments in the domain. Given such a diagram, we can effectively calculate the denotation in $\mathscr{I}$ of any term in Γ, then calculate the truth-value in $\mathscr{I}$ of any atomic sentence of which any sentence of Γ is compounded, and then, via the rules of interpretation for the propositional connectives (cases 1–5 of Chapter 9, i.e. essentially, via the usual truth tables) calculate the truth-values which $\mathscr{I}$ assigns to the several sentences which make up Γ.

Then our effective procedure is this: write down all diagrams of all interpretations of Γ with domain $\{1, \ldots, m\}$ $(1 \leqslant m \leqslant n)$, calculate the truth-values of the members of Γ in each of them, and see whether all members of Γ are true in at least one of these interpretations. If so, Γ is satisfiable; if not, not.

Exercises

12.1 Use the construction given in the proof of Lemma III to find interpretations which match the sets of quantifier-free sentences in canonical derivations from the following sets:

(*a*) $\{\forall x \exists y\, x\text{Ł}y\}$; (*b*) $\{\exists x\, x\text{Ł}f(gx, x)\}$.

12.2 A set Γ of sentences is said to have *arbitrarily large finite models* when, for each positive integer n, there is a model of Γ whose domain is finite and has at least n members. Show that

> If Γ has arbitrarily large finite models, it has a model whose domain is infinite.

Conclude that there is no *axiom of finitude*, i.e., no sentence which is true in all and only those interpretations whose domains are finite.

12.3 Show that there is an effective procedure for deciding the validity of prenex sentences in which no function symbols occur, and in which no existential quantifier is to the left of any universal quantifier.

12.4 Show ('from scratch') that if Γ is a set of sentences which contain only sentence letters, connectives, and parentheses, then Γ is satisfiable if every finite subset is. (*Hint*: strip from the proof of Lemma III the numerous complexities which are required when terms, '=' and other predicate letters may occur in Γ.)

Solutions

12.1(*a*) $\Delta = \{\forall x \exists y\, x\text{Ł}y\}$. A canonical derivation $\mathscr{D}$ from Δ:

1	$\forall x \exists y\, x\text{Ł}y$	Δ
2	$\exists y\, a_1\text{Ł}y$	1
3	$a_1\text{Ł}a_2$	2
4	$\exists y\, a_2\text{Ł}y$	1
5	$a_2\text{Ł}a_3$	4
	$\vdots$	
2*n*	$\exists y\, a_n\text{Ł}y$	1
2*n*+1	$a_n\text{Ł}a_{n+1}$	2*n*

Here $\Gamma = \{a_1\text{Ł}a_2, a_2\text{Ł}a_3, \ldots, a_n\text{Ł}a_{n+1}, \ldots\}$. There is some arbitrariness in the choice of the sequence $A_1, A_2, \ldots$, but once that sequence is fixed (e.g. as at the left, below), the sequences Γ_1, $\Gamma_2, \ldots$, and $B_1, B_2, \ldots$ are fixed and thereby $\mathscr{I}$ is determined.

$A_1: a_1La_1$	$\Gamma_2 = \{a_1La_1\} \cup \Gamma$	$B_1: a_1La_1$
$A_2: a_1 = a_1$	$\Gamma_3 = \Gamma_2 \cup \{a_1 = a_1\}$	$B_2: a_1 = a_1$
$A_3: a_1La_2$	$\Gamma_4 = \Gamma_3$	$B_3: a_1\text{Ł}a_2$
$A_4: a_1 = a_2$	$\Gamma_5 = \{a_1 \neq a_2\} \cup \Gamma_4$	$B_4: a_1 \neq a_2$
$A_5: a_2La_1$	$\Gamma_6 = \{a_2La_1\} \cup \Gamma_5$	$B_5: a_2La_1$
$A_6: a_2 = a_1$	$\Gamma_7 = \Gamma_6 \cup \{a_2 \neq a_1\}$	$B_6: a_2 \neq a_1$

If the subscripts on the pairs of names flanking '*L*' and '=' in the sequence of *A*s continue in the pattern

22, 13, 31, 23, 32, 33, 14, 41, 24, 42, 34, 43, 44, ...

then the sequence of *B*s continues:

$$a_2La_2,\ a_2 = a_2,\ a_1La_3,\ a_1 \neq a_3,\ a_3La_1,\ a_3 \neq a_1,$$
$$a_2\text{Ł}a_3,\ a_2 \neq a_3,\ a_3La_2,\ a_3 \neq a_2,\ a_3La_3,\ a_3 = a_3,$$
$$a_1La_4,\ a_1 \neq a_4, \ldots$$

where distinct names are always asserted to name distinct objects in the domain of $\mathscr{I}$ – i.e., $[a_m] \neq [a_n]$ if and only if $m \neq n$ – and where the relation assigned to '*L*' by $\mathscr{I}$ is asserted to hold between all objects except those between which it is asserted not to hold in Γ, i.e., $\mathscr{I}(a_mLa_n) = 0$ if $n = m+1$ and otherwise $\mathscr{I}(a_mLa_n) = 1$. Had the sequence of *A*s been chosen differently, the matching interpretation might have been very different.

12.2 For each $i \geqslant 2$, let A_i be a sentence which is true in an interpretation if and only if there are at least i members of the domain of the interpretation. E.g. A_3 might be

$$\exists x_1 \exists x_2 \exists x_3 (x_1 \neq x_2 \,\&\, x_2 \neq x_3 \,\&\, x_1 \neq x_3).$$

Let Γ' be the set of all such sentences, $\Gamma' = \{A_1, A_2, \ldots\}$. Every finite subset of $\Gamma \cup \Gamma'$ has a model. (Why?) So by the compactness theorem, $\Gamma \cup \Gamma'$ has a model. No model of Γ' can have a finite domain, so $\Gamma \cup \Gamma'$ and hence Γ have a model with infinite domain.

13 The Skolem–Löwenheim theorem

The cluster of theorems that is generally called 'the Skolem–Löwenheim theorem' is a group of results that concern the size of domains of interpretations of sentences in first-order logical languages, 'size' being understood in the sense of (*cardinal*) *number of members*. They typically have the form: if there is an interpretation with (semantical) property —, then there is an interpretation with (semantical) property —, whose domain has size – – –. For example, the best known Skolem–Löwenheim theorem, an easy consequence of the version proved in Chapter 12, reads: if there is an interpretation in which a sentence S is true, then there is an interpretation in which S is true, whose domain is enumerable (Löwenheim, 1915). In the present chapter we shall prove a strong form of the Skolem–Löwenheim theorem from which this and other versions follow.

In Chapter 25 we shall prove that if S is a sentence that contains k *one-place* predicate letters and r variables, and possibly also the equals-sign '=', but no names, function symbols, or two or more place predicate letters (a so-called 'monadic' sentence), then if S is true in any interpretation at all, it is true in one whose domain contains no more than $2^k \cdot r$ members. Thus if S is a monadic sentence true in some interpretation whose domain is infinite, S is also true in some other interpretation whose domain is finite. One might wonder whether the restriction to monadic sentences is essential. It is.

Let (a) = '$\forall x - xRx$', let (b) = '$\forall x \exists y\, xRy$', and let

$$(c) = \text{‘}\forall x \forall y \forall z ((xRy \,\&\, yRz) \rightarrow xRz)\text{’}.$$

Let (1) be the conjunction of (a), (b), and (c). (1) is then a sentence which is true in no interpretation with a finite domain, but is true in some interpretation with an infinite domain.

To see that (1) is true in no interpretation with a finite domain, suppose that $\mathscr{I}((1)) = 1$, that D = the domain of $\mathscr{I}$, and that $\mathscr{I}$ specifies that 'R' is to be true of c, d iff cSd. Let's call a sequence $d_1, \ldots, d_n$ of elements of D a *good* sequence if d_iSd_j whenever $i < j \leqslant n$. Observe that if $d_1, \ldots, d_n$ is a good sequence, then $d_1, \ldots,$ and d_n are all *distinct*, for if $i < j$, but $d_i = d_j$, then d_iSd_i, which is impossible, as $\mathscr{I}((a)) = 1$. We shall show that D is infinite by showing inductively that for each positive n, there is a good sequence $d_1, \ldots, d_n$. Since D is nonempty, there is, trivially, a good sequence containing just one member d_1 of D. Suppose that $d_1, \ldots, d_n$ is a good sequence. Since $\mathscr{I}((b)) = 1$, for some d_{n+1} in D, d_nSd_{n+1}. But then $d_1, \ldots, d_n, d_{n+1}$ is also a good sequence, for if $i < n$, then, since d_iSd_n, d_nSd_{n+1}, and $\mathscr{I}((c)) = 1$, we have that d_iSd_{n+1}, and therefore we have that if either $i < j \leqslant n$ or $i < j = n+1$, then d_iSd_j, and thus have that d_iSd_j whenever $i < j \leqslant n+1$.

On the other hand, (*a*) no real number is less than itself; (*b*) every real number is less than some other real number; and (*c*) if one real is less than a second, which is less than a third, the first is less than the third. (1) is therefore true in the interpretation $\mathscr{I}$ whose domain is the set of all real numbers and which specifies that 'R' is to be true of c, d iff $c < d$. The domain of this interpretation is unnecessarily large, however, for (1) is also true in the interpretation $\mathscr{J}$ whose domain is the set of all natural numbers, and which, like $\mathscr{I}$, specifies that 'R' is to be true of c, d iff $c < d$. The domain of $\mathscr{I}$ is non-enumerable, that of $\mathscr{J}$, enumerable. Another question thus suggests itself: is there any sentence of a first-order logical language which, though true in some interpretation, is true only in interpretations whose domains are non-enumerable?

Löwenheim's 1915 theorem, stated five paragraphs back, answers this question negatively. In 1919 Skolem extended Löwenheim's result in a far-reaching way by proving a theorem that immediately implies not only Löwenheim's theorem, but the stronger statement proved in Chapter 12 that if an enumerable collection of sentences is satisfiable, then there is a single interpretation with an enumerable domain in which all members of the collection are true. Skolem's 1919 theorem was that *any interpretation*† *has an elementarily equivalent subinterpretation with an enumerable domain.*

What are 'subinterpretations' and what's 'elementary equivalence'?

An interpretation $\mathscr{J}$ is called a *subinterpretation* of an interpretation $\mathscr{I}$ if $\mathscr{J}$ and $\mathscr{I}$ are interpretations of the same languages and

(1) The domain E of $\mathscr{J}$ is a subset of the domain of $\mathscr{I}$;
(2) $\mathscr{J}$ assigns names the same designations as $\mathscr{I}$;
(3) $\mathscr{J}$ assigns an n-place function symbol f the function $\mathfrak{g}$ which takes values only on sequences $c_1, \ldots, c_n$ of objects in E and for which

$$\mathfrak{g}(c_1, \ldots, c_n) = \mathfrak{f}(c_1, \ldots, c_n),$$

where $\mathfrak{f}$ is the function $\mathscr{I}$ assigns f;

(4) $\mathscr{J}$ assigns sentence-letters the same truth-values as $\mathscr{I}$; and
(5) $\mathscr{J}$ specifies that an n-place predicate letter R is to be true of a sequence $c_1, \ldots, c_n$ of objects in E iff $\mathscr{I}$ specifies that R is to be true of $c_1, \ldots, c_n$.

Which sets E may be domains of subinterpretations of $\mathscr{I}$? First of all, E must be a nonempty subset of the domain of $\mathscr{I}$. Clause 2 imposes another requirement on E: that it contain all designations that $\mathscr{I}$ assigns names. Clause 3 imposes a further requirement: that if $c_1, \ldots, c_n$ are in E, and $\mathscr{I}$ assigns the function $\mathfrak{f}$ to some n-place function symbol, then $\mathfrak{f}(c_1, \ldots, c_n)$ must also be in E. If these three requirements are met, however, there is a unique subinterpretation of $\mathscr{I}$ whose domain is E. If nonempty, the set of denotations that $\mathscr{I}$ assigns to terms is a set that meets all three requirements.

If $\mathscr{J}$ is a subinterpretation of $\mathscr{I}$, $\mathscr{J}$ interprets the same sentences as $\mathscr{I}$: the same sentences have truth-values in both, though not necessarily the same truth-values. It follows from clauses 2 and 3 that $\mathscr{J}$ assigns each term the same denotation as $\mathscr{I}$. It then follows from clauses 4 and 5 that atomic sentences have the same truth-values in both, and therefore so do all *quantifier-free* sentences.

Two interpretations $\mathscr{I}$ and $\mathscr{J}$ are said to be *elementarily equivalent* if they interpret the same sentences and any sentence is true in $\mathscr{I}$ iff it is true in $\mathscr{J}$. This definition can be weakened to: they interpret the same sentences and all sentences true in $\mathscr{I}$ are true in $\mathscr{J}$; for if $\mathscr{I}$ and $\mathscr{J}$ interpret S and S is true in $\mathscr{J}$, then S's negation is not true in $\mathscr{J}$, hence not true in $\mathscr{I}$, and thus S is true in $\mathscr{I}$.

Skolem showed that for any interpretation $\mathscr{I}$, there is an interpretation $\mathscr{J}$ which (A) is a subinterpretation of $\mathscr{I}$, (B) has an enumerable domain, and (C) is elementarily equivalent to $\mathscr{I}$. It follows that given an interpretation of all the (enumerably many) sentences in a (first-order) language, no matter how large its domain, one can whittle down the domain in such a way that even though only an enumerable number of elements remains, the very same sentences are true in the reduced interpretation as were true in the original.

There are related results – 'upward' theorems – that speak of arbitrarily large models, and still stronger 'downward' theorems than the one we shall demonstrate, but we shall not discuss them (the reader is referred

† We assumed in Chapter 9 that no interpretation assigns appropriate objects to non-enumerably many non-logical symbols, and hence that every interpretation interprets (defines truth-values for) at most enumerably many sentences. This assumption is vital to the proof of the Skolem–Löwenheim theorem (in Skolem's 1919 version) (cf. Exercise 13.1), as is the restriction to first-order logic (cf. Chapter 17).

to Chang & Keisler's *Model Theory*). We shall now prove Skolem's 1919 theorem and derive another Skolem–Löwenheim theorem as a corollary. The theorem follows quite directly from the strong soundness theorem of Chapter 11, Lemmas I and II of Chapter 12, and the facts about subinterpretations just mentioned.

Theorem

Any interpretation $\mathscr{I}$ has an elementarily equivalent subinterpretation $\mathscr{J}$ whose domain is enumerable.

Proof. Let Δ be the set of all (prenex) sentences true in $\mathscr{I}$. By Lemma I of Chapter 12, there is a canonical derivation $\mathscr{D}$ from Δ. If a is a name or f is a function symbol assigned a denotation or a function by $\mathscr{I}$, then all of $a = a$, $\forall x f(x, \ldots, x) = f(x, \ldots, x)$, and $\forall x\, x = x$ occur in Δ and $\mathscr{D}$. By clause 5 of the definition of 'canonical derivation', all function symbols appearing in $\mathscr{D}$ appear in Δ, and so by the strong soundness theorem of Chapter 11, there is an interpretation $\mathscr{L}$, in which all sentences in $\mathscr{D}$ are true, and which differs from $\mathscr{I}$ only in what it assigns to those names that appear in $\mathscr{D}$ but are assigned no denotation by $\mathscr{I}$.

Let Γ = the set of quantifier-free sentences in $\mathscr{D}$. As $\mathscr{L}$ is a model of Γ, any term t appearing in Γ is assigned a denotation by $\mathscr{L}$; and if t is assigned a denotation by $\mathscr{L}$, t is formed from names and function symbols assigned things by $\mathscr{I}$, all of which appear in $\mathscr{D}$, and other names appearing in $\mathscr{D}$, and hence, as $\forall x\, x = x$ is in $\mathscr{D}$, $t = t$ is in Γ. Thus the terms assigned denotations by $\mathscr{L}$ are just those appearing in Γ.

Let E be the set of denotations $\mathscr{L}$ assigns to terms. E is non-empty, as $t = t$ is in Γ (for some term t). E is enumerable, because Γ is. Let $\mathscr{K}$ be the subinterpretation of $\mathscr{L}$ with domain E. $\mathscr{K}$ assigns a term the denotation d iff $\mathscr{L}$ assigns it d. Every object in $\mathscr{K}$'s domain is therefore denoted (according to $\mathscr{K}$) by a term appearing in Γ. All members of Γ, being quantifier-free and true in $\mathscr{L}$, are true in $\mathscr{L}$'s subinterpretation $\mathscr{K}$. So $\mathscr{K}$ matches Γ (cf. Chapter 12). So by Lemma II of Chapter 12, $\mathscr{K}$ is a model of the set of all sentences in $\mathscr{D}$, and hence of Δ.

If $\mathscr{J}$ is just like $\mathscr{K}$ except that it assigns denotations to the same names as $\mathscr{I}$, then $\mathscr{J}$ is a model of Δ with an enumerable domain, and thus an elementarily equivalent subinterpretation of $\mathscr{I}$ with an enumerable domain.

Corollary 1 (Skolem, 1919)

If θ is an enumerable collection of sentences that is satisfiable, there is an interpretation $\mathscr{J}$ with enumerable domain which is a model of θ.

Corollary 2 (Löwenheim, 1915)

If a sentence is satisfiable, then it is true in some interpretation with an enumerable domain.

Corollary 3

If a sentence that does not contain the equals-sign '=' is satisfiable, then it is true in some interpretation whose domain is the set of all natural numbers.

Proof. This corollary follows from Corollary 2 together with these two facts: (a) If a sentence not containing '=' is true in some interpretation $\mathscr{J}$ with finite domain, it is true in some interpretation $\mathscr{I}$ with enumerably infinite domain. (b) If a sentence is true in some interpretation $\mathscr{I}$ with enumerably infinite domain, it is true in some interpretation $\mathscr{K}$ whose domain is the set of natural numbers.

Proof of (a). Suppose that $\mathscr{J}$ is an interpretation with finite domain E. Let e be some element of E. Let $c_0, c_1, \ldots$ be an enumerably infinite sequence of objects *not* in E. Let D be the set that contains all members of E and all the c_is. If d is in D, define d^* as follows: $d^* = d$ if d is in E, and $d^* = e$ if d is not in E. Let $\mathscr{I}$ be the following interpretation, 'which makes all the c_is indistinguishable from e': The domain of $\mathscr{I}$ is D. $\mathscr{I}$ assigns names and sentence letters whatever $\mathscr{J}$ assigns them; $\mathscr{I}$ specifies that an n-place predicate letter R is to be true of $d_1, \ldots, d_n$ iff $\mathscr{J}$ specifies that R is to be true of $d_1^*, \ldots, d_n^*$; and $\mathscr{I}$ assigns an n-place function symbol f the function $\mathfrak{f}$ such that for any $d_1, \ldots, d_n$ in D, $\mathfrak{f}(d_1, \ldots, d_n) = \mathfrak{g}(d_1^*, \ldots, d_n^*)$, where $\mathfrak{g}$ is the function $\mathscr{J}$ assigns f. We shall show that for any sentences S not containing '=', $\mathscr{I}(S) = \mathscr{J}(S)$.

To that end, for each d in D, choose a *new name* a_d, i.e., one not assigned any designation by $\mathscr{J}$. (Choose different names for different members of D.) Let $\mathscr{I}_1$ be just like $\mathscr{I}$ except that for each a_d, $\mathscr{I}_1(a_d) = d$; let $\mathscr{J}_1$ be just like $\mathscr{J}$ except that for each a_d, $\mathscr{J}_1(a_d) = d^*$. Then for *any* name b, if $\mathscr{I}_1(b) = d$, $\mathscr{J}_1(b) = d^*$. If now, for any i between 1 and n, $\mathscr{I}_1(t_i) = d_i$ and $\mathscr{J}_1(t_i) = d_i^*$, then

$$\mathscr{I}_1(f(t_1, \ldots, t_n)) = \mathfrak{f}(d_1, \ldots, d_n) = \mathfrak{g}(d_1^*, \ldots, d_n^*) = \mathscr{J}_1(f(t_1, \ldots, t_n)).$$

It follows inductively that for any term t, if $\mathscr{I}_1(t) = d$, $\mathscr{J}_1(t) = d^*$. Hence $\mathscr{I}_1(Rt_1 \dots t_n) = 1$ iff $\mathscr{J}_1(Rt_1 \dots t_n) = 1$. Thus every atomic sentence not containing '$=$' is true in $\mathscr{I}_1$ iff true in $\mathscr{J}_1$. If we can conclude that *every* '$=$'-free sentence has the same truth-value in $\mathscr{I}_1$ as in $\mathscr{J}_1$, we can conclude that the same holds good of $\mathscr{I}$ and $\mathscr{J}$; for a sentence not containing any of the a_ds will be true in $\mathscr{I}$ iff true in $\mathscr{I}_1$, and true in $\mathscr{J}$ iff true in $\mathscr{J}_1$. As a truth-functional compound of sentences having the same truth-values in $\mathscr{I}_1$ as in $\mathscr{J}_1$ clearly has the same truth-value in $\mathscr{I}_1$ as in $\mathscr{J}_1$, if there is a simplest sentence S having opposite truth-values in $\mathscr{I}_1$ and $\mathscr{J}_1$, S must be of one of the forms $\exists v\, H$ and $\forall v\, H$. Suppose $S = \exists v\, H$. (The argument is similar if $S = \forall v\, H$.) But then, since any member of D or E is the denotation according to $\mathscr{I}_1$ or $\mathscr{J}_1$, respectively, of some name, we have that $\mathscr{I}_1(S) = 1$ iff $\mathscr{I}_1(\exists v\, H) = 1$, iff for some name a, $\mathscr{I}_1(H_v\, a) = 1$, iff – as $H_v\, a$ is *simpler* than $\exists v\, H$ – for some name a, $\mathscr{J}_1(H_v\, a) = 1$, iff $\mathscr{J}_1(\exists v\, H) = 1$, iff $\mathscr{J}_1(S) = 1$. Thus no '$=$'-free sentence S has opposite truth-values in $\mathscr{I}_1$ and $\mathscr{J}_1$, and therefore the same goes for $\mathscr{I}$ and $\mathscr{J}$. So (a) holds.

The assumption that S not contain '$=$' is indispensable: '$\exists x \forall y\, x = y$' is true in any interpretation whose domain contains exactly one member, but in no interpretation with a larger domain.

Proof of (b). Let $d_0, d_1, d_2, \dots$ be a repetition-free enumeration of all members of the domain of $\mathscr{I}$. Then the same sentences are true in $\mathscr{I}$ as in the interpretation $\mathscr{K}$ whose domain is the set of all natural numbers, and which assigns each sentence letter the same truth-value as $\mathscr{I}$, assigns a name the number i as designation iff $\mathscr{I}$ assigns it d_i, specifies that a predicate letter is to be true of a sequence of natural numbers $i_1, i_2, \dots, i_n$ iff $\mathscr{I}$ specifies that it is to be true of $d_{i_1}, d_{i_2}, \dots, d_{i_n}$, and assigns a function symbol f the function from the one $\mathscr{I}$ assigns f by similarly everywhere 'replacing' d_{i_j} by i_j.

At one time the Skolem–Löwenheim theorem was considered philosophically perplexing because some of its consequences were perceived as anomalous. The apparent anomaly, sometimes called 'Skolem's paradox', is that there exist certain interpretations in which a certain sentence, which seems to say that non-enumerably many sets of natural numbers exist, is true, even though the domains of these interpretations contain only enumerably many sets of natural numbers, and the predicate letter in the sentence we would be inclined to translate as 'set (of natural numbers)' is true of just the sets (of natural numbers) in the domains.

There is no denying that the state of affairs thought to be paradoxical does obtain. In order to see how it arises, we shall first need an alternative account of what it is for a set E of sets of natural numbers to be enumerable, and for this we shall need a few definitions.

We shall say that one (ordered) pair, $\langle x, y \rangle$ of natural numbers *precedes* another $\langle i, j \rangle$ *in order* O if either $x+y < i+j$ or $(x+y = i+j$ and $x < i)$. We define the *pairing function* J by setting $J(x,y) = z$ if $\langle x, y \rangle$ is the $(z+1)$st pair in order O. J is then a one–one function from the sets of pairs of natural numbers onto the set of natural numbers, i.e. J assigns each pair of natural numbers as argument a unique natural number as value, assigns different pairs different numbers, and assigns every number to some pair or other. (J, incidentally, is recursive.)

We shall call a set w of natural numbers an *enumerator* of a set E of sets of natural numbers if

$$\begin{aligned}&\forall z\,(z \text{ is a set of natural numbers } \& \ z \text{ is in } E \rightarrow\\ &\exists x\,(x \text{ is a natural number } \& \ \forall y\,(y \text{ is a natural number} \rightarrow\\ &(y \text{ is in } z \leftrightarrow J(x,y) \text{ is in } w)))).\end{aligned}$$

The fact about enumerators and enumerability that we need is that *a set E of sets of natural numbers is enumerable iff E has an enumerator*.

(The reason: suppose E is enumerable. Let $e_0, e_1, e_2, \dots$ be an enumeration of sets of natural numbers that contains all of the members of E, and possibly some other sets of natural numbers too. Then the set of numbers $J(x,y)$ such that y is in e_x is an enumerator of E. Conversely, if w is an enumerator of E, then, where $e_x =$ the set of those numbers y such that $J(x,y)$ is in w, $e_0, e_1, e_2, \dots$ is an enumeration that contains all members of E, and therefore E is enumerable.)

We want now to look at a language and some of its interpretations. The language contains just the names '**0**', '**1**', '**2**', etc., one two-place function symbol '**J**', two one-place predicate letters '**N**' and '**S**', and one two-place predicate letter '**ϵ**'. The interpretations we are interested in are those whose domains contain all natural numbers and some (possibly all) sets of natural numbers, but nothing else; which assign each numeral its ordinary designation; which assign '**J**' the pairing function (extended so as to take some arbitrary value – say 17 – for the argument x, y if either x or y is a set); and which specify that '**N**' is to be true of any number, '**S**' is to be true of any set of numbers, and '**ϵ**' is to be true of y, z iff the number y is in the set z. One of these interpretations, $\mathscr{I}$, is called the *standard* interpretation. It is the one whose domain contains *all* sets of natural numbers.

In all of these interpretations the sentence

$$-\exists w(\mathrm{S}w \,\&\, \forall z(\mathrm{S}z \to \exists x(\mathrm{N}x \,\&\, \forall y(\mathrm{N}y \to (y \in z \leftrightarrow \mathrm{J}(x,y) \in w))))) \qquad (2)$$

will have a truth-value. It will be *true* in one of them iff there is no enumerator of the set of all sets of numbers in the domain of the interpretation *that is itself in the domain of the interpretation* (as we can see by checking back to the definition of 'enumerator'). We can't simply say that the sentence is true in an interpretation iff there is no enumerator of the set of all sets of numbers in the domain, because the quantifier '$\exists w$' is understood to range over, or 'refer to', members of the domain of the interpretation alone.

There is, as we know, *no* enumerator of the set of all sets of numbers in the domain of $\mathscr{I}$ since all sets of numbers are in $\mathscr{I}$'s domain. *A fortiori*, there is no such enumerator in the domain of $\mathscr{I}$, and sentence (2) is therefore true in $\mathscr{I}$, and can be said to mean 'Non-enumerably many sets exist', when interpreted 'over' $\mathscr{I}$, since it then denies that there is an enumerator of the set of all sets of numbers. By the Skolem–Löwenheim theorem, $\mathscr{I}$ has an elementarily equivalent subinterpretation $\mathscr{J}$, whose domain is enumerable and thus contains only enumerably many sets of numbers. (All of 0, 1, 2, etc. are in $\mathscr{J}$'s domain since these are the designations $\mathscr{I}$, and hence $\mathscr{J}$, assigns to '**0**', '**1**', '**2**', etc.) Since $\mathscr{I}$ and $\mathscr{J}$ are elementarily equivalent, (2) is a sentence true in $\mathscr{J}$, and therefore in an interpretation in whose domain there are only enumerably many sets of numbers, and in which 'S' is true of just the sets of numbers in its domain. This is Skolem's 'paradox'.

How is the paradox to be resolved? Well, although the set of all sets in the domain of $\mathscr{J}$ does indeed have an enumerator, since it is enumerable, none of its enumerators can be *in* the domain of $\mathscr{J}$ (for otherwise,

$$(\mathrm{S}w \,\&\, \forall z(\mathrm{S}z \to \exists x(\mathrm{N}x \,\&\, \forall y(\mathrm{N}y \to (y \in z \leftrightarrow \mathrm{J}(x,y) \in w)))))$$

would be true in $\mathscr{J}$ of one of them and thus (2) would be false in $\mathscr{J}$.) So part of the explanation of how (2) can be true in $\mathscr{J}$ is that those sets which 'verify' the claim that the set of sets in the domain of $\mathscr{J}$ is enumerable are not themselves members of the domain of $\mathscr{J}$.

A further part of the explanation is that what a sentence should be understood as saying or meaning or denying is at least as much a function of the interpretation over which the sentence is interpreted (and even of the way in which that interpretation is described or referred to) as of the symbols that constitute it. (2) can be understood as saying 'non-enumerably many sets exist' when its quantifiers are understood as ranging over a collection containing all numbers and all sets of numbers, such as the domain of the standard interpretation $\mathscr{I}$, but it cannot be so understood when its quantifiers range over other domains, in particular, not when they range over members of countable domains. The sentence (2) – that sequence of symbols – 'says' something only when supplied with an interpretation. It may be surprising and even amusing that (2) is true in all sorts of interpretations, including, perhaps, some subinterpretations $\mathscr{J}$ of $\mathscr{I}$ that have enumerable domains, but it should not *a priori* seem impossible that it be true in these. Interpreted over such a $\mathscr{J}$, it will only say 'the domain of $\mathscr{J}$ contains no enumerator of the set of sets of numbers in $\mathscr{J}$' which is, of course, true.

Exercises

13.1 Suppose that interpretations were allowed to assign objects to non-enumerably many one-place predicate letters. Show that the Skolem–Löwenheim theorem, as we stated it, would then be false.

13.2 A subinterpretation $\mathscr{J}$ of $\mathscr{I}$ is called an *elementary subinterpretation* of $\mathscr{I}$ if for every formula F of the language of $\mathscr{I}$ and every sequence $o_1, \ldots, o_n$ in the domain of $\mathscr{J}$

$$\mathscr{I}^{a_1 \ldots a_n}_{o_1 \ldots o_n}(F^*) = \mathscr{J}^{a_1 \ldots a_n}_{o_1 \ldots o_n}(F^*).$$

Here F is supposed to contain at most the n variables $v_1, \ldots, v_n$ free, $a_1, \ldots, a_n$ are n names to which $\mathscr{I}$ assigns no designation, and F^* is the result of substituting $a_1, \ldots, a_n$ for all free occurrences of $v_1, \ldots, v_n$ (respectively) in F. Show that any interpretation $\mathscr{I}$ has an elementary subinterpretation $\mathscr{J}$ whose domain is enumerable; deduce the version of the Skolem–Löwenheim theorem proved in the text from this statement.

13.3 Show that the version of the Skolem–Löwenheim theorem proved in the text implies the axiom of dependent choice. (Cf. Exercise 11.5.)

Solutions

13.1 Let $\mathbb{R}$ be the set of real numbers. For each r in $\mathbb{R}$ let A_r be a one-place predicate letter. Let $\mathscr{I}$ be the interpretation with domain $\mathbb{R}$ which specifies that each A_r is to be true of r and r alone. For each r, $\exists x A_r x$ is true in $\mathscr{I}$, and hence true in every elementarily equivalent subinterpretation. Each real number r must therefore belong to the domain of every elementarily equivalent subinterpretation $\mathscr{J}$. Each such $\mathscr{J}$ will therefore have a non-enumerable domain.

13.3 Suppose X is a nonempty set, and for any x in X there is a y in X such that xRy. Let $\mathscr{I}$ be the interpretation whose domain is X, and according to which '**R**' is true of x, y iff xRy. $\forall x \exists y\, x\mathbf{R}y$ is true in $\mathscr{I}$. Let $\mathscr{J}$ be an elementarily equivalent subinterpretation of $\mathscr{I}$ whose domain E is enumerable. Let $e_0, e_1, e_2, \ldots$ be an enumeration of E. $\forall x \exists y\, x\mathbf{R}y$ is true in $\mathscr{J}$. Therefore for every e_i in E there will be an e_j in E such that $e_i R e_j$. Define f by: $f(0) = e_0$; for each n, $f(n+1) = e_j$ iff $f(n)\ Re_j$ and for every $k < j$, not: $f(n)\ Re_k$. The axiom of dependent choice is not required to guarantee the existence of f.

14
Representability in Q

The present chapter falls into three parts. In the first part we introduce the notion of *representability of a function* (of natural numbers) *in a theory* and present a theory, called 'Q'. In the second part we give an alternative characterization of the recursive functions,[†] and in the third we use this new characterization to show that every recursive function is representable in the theory Q. In the next chapter several important results about undecidability, indefinability and incompleteness will be shown to follow from the latter result. The converse, that every function representable in Q is recursive, is also true, and we shall also indicate why at the end of the next chapter (Exercise 15.2).

Part I

We shall take a theory to be a set of sentences in some language that contains all of its logical consequences that are sentences in that language. If a sentence A is a member of theory T, it is called a *theorem* of T; to indicate that A is a theorem of T, we write: $\vdash_T A$.

From now through Chapter 21, we shall confine our attention to *numerical* theories: theories whose language contains the name **o** and the one-place function symbol $'$. (Q will be such a theory.) The *numeral* for n, **n**, is the result of attaching n occurrences of $'$ to (the right of) **o**. Thus $\mathbf{3} = \mathbf{o}'''$ and the numeral for $n+1$ is $\mathbf{n}'$. For any natural number n, **n** is an expression or sequence of symbols, a *term* of the sort described.

If A is a formula that contains free occurrences of the n (distinct) variables $x_1, \ldots, x_n$, we shall sometimes refer to A as $A(x_1, \ldots, x_n)$. For any natural numbers $p_1, \ldots, p_n$, $A(\mathbf{p}_1, \ldots, \mathbf{p}_n)$ is the result of substituting an occurrence of $\mathbf{p}_i$ for each free occurrence of x_i in $A(x_1, \ldots, x_n)$ (for each i between 1 and n). In discussing a formula $A(x_1, \ldots, x_n)$ we may wish to consider a formula, which we refer to as '$A(y_1, \ldots, y_n)$'. This is to be understood to be the formula that results when any bound occurrence of y_i that may occur in $A(x_1, \ldots, x_n)$ is first replaced by an occurrence of a new

[†] Cf. the last paragraph of Chapter 8.

variable z_i (different z_is for different y_is), and then an occurrence of y_i is substituted for each free occurrence of x_i in the result.

To reduce clutter, we shall write 'p' instead of '$p_1, \ldots, p_n$', 'x' instead of '$x_1, \ldots, x_n$', and '$\mathbf{p}$' instead of '$\mathbf{p_1}, \ldots, \mathbf{p_n}$'.

We can now define representability. An n-place function f is *representable* in a theory T if there is a formula $A(\mathsf{x}, x_{n+1})$ such that for any natural numbers p, j, if $f(\mathsf{p}) = j$, then $\vdash_T \forall x_{n+1}(A(\mathbf{p}, x_{n+1}) \leftrightarrow x_{n+1} = \mathbf{j})$. In this case the formula $A(\mathsf{x}, x_{n+1})$ is said to *represent* f in T.

The requirement that $\vdash_T \forall x_{n+1}(A(\mathbf{p}, x_{n+1}) \leftrightarrow x_{n+1} = \mathbf{j})$ should hold whenever $f(\mathsf{p}) = j$ is equivalent to the requirement that both $\vdash_T A(\mathbf{p}, \mathbf{j})$ and $\vdash_T \forall x_{n+1}(A(\mathbf{p}, x_{n+1}) \rightarrow x_{n+1} = \mathbf{j})$ should hold whenever $f(\mathsf{p}) = j$. If the sentence $\mathbf{j} \neq \mathbf{k}$ is a theorem of T whenever $j \neq k$ (and we shall see that Q is a theory of which this is so), then if A represents f in T and $f(\mathsf{p}) \neq k$, then $\vdash_T -A(\mathbf{p}, \mathbf{k})$ (for $\vdash_T \mathbf{j} \neq \mathbf{k}$, where $j = f(\mathsf{p})$).

The language of theory Q is L, the *language of arithmetic*. L contains four non-logical symbols, the name $\mathbf{o}$, the one-place function symbol $'$, and two two-place function symbols, $+$ and $\cdot$. Q is the set of sentences in L that are logical consequences of these seven sentences, the *axioms of Q*:

Q 1 $\forall x \forall y(x' = y' \rightarrow x = y)$,

Q 2 $\forall x\, \mathbf{o} \neq x'$,

Q 3 $\forall x (x \neq \mathbf{o} \rightarrow \exists y\, x = y')$,

Q 4 $\forall x\, x + \mathbf{o} = x$,

Q 5 $\forall x \forall y\, x + y' = (x+y)'$,

Q 6 $\forall x\, x \cdot \mathbf{o} = \mathbf{o}$,

Q 7 $\forall x \forall y\, x \cdot y' = (x \cdot y) + x$.

Q is a consistent theory, for all of its axioms are true in the *standard interpretation $\mathcal{N}$ for its language L*, in which the domain is the set of all natural numbers, $\mathbf{o}$ is assigned zero as denotation, and $'$, $+$, and $\cdot$ are assigned the successor, addition, and multiplication functions. Q is a theory that is rather strong in certain ways (all recursive functions are representable in it), but rather weak in others (e.g. $\forall x \forall y\, x + y = y + x$ is not a theorem of Q, as an exercise at the end of the chapter shows). Tarski, Mostowski, and R. Robinson have written that it 'is distinguished by the simplicity and clear mathematical content of its axioms'. We shall devote the remainder of this chapter to showing that all recursive functions are representable in Q.

Part II

We recall from Chapters 7 and 8 that the recursive functions can be characterized as those functions obtainable from the zero function, the successor function and the identity functions by means of a finite number of applications of the operations of composition, primitive recursion, and minimization of those functions called *regular* functions.

The *zero* function z is the one-place function whose value for all arguments is zero.

The *successor* function $'$ $(= s)$ is the one-place function whose value for any argument i is $i + 1$ $(= i'$, the successor of $i)$.

For each $m \geqslant 1$ and each $n \leqslant m$, there is an m-place *identity* function id^m_n. For any natural numbers $i_1, \ldots, i_m$, $\mathrm{id}^m_n(i_1, \ldots, i_m) = i_n$.

If f is an m-place function, and $g_1, \ldots, g_m$ are all n-place functions, then the n-place function h is said to be obtained from $f, g_1, \ldots, g_m$ by *composition* if for any natural numbers p, $h(\mathsf{p}) = f(g_1(\mathsf{p}), \ldots, g_m(\mathsf{p}))$.

If f is an n-place function and g is an $(n+2)$-place function, then the $(n+1)$-place function h is said to be obtained from f and g by *primitive recursion* if for any natural numbers p, k, $h(\mathsf{p}, 0) = f(\mathsf{p})$ and

$$h(\mathsf{p}, k+1) = g(\mathsf{p}, k, h(\mathsf{p}, k)).$$

An $(n+1)$-place function f is called *regular* if for any natural numbers p, there exists at least one natural number i such that $f(\mathsf{p}, i) = 0$. If f is a regular $(n+1)$-place function, then the n-place function g is said to be obtained from f by *minimization* if for any natural numbers p,

$$g(\mathsf{p}) = \mu i f(\mathsf{p}, i) = 0,$$

where 'μi' means 'the least natural number i such that'.

If R is an n-place relation of natural numbers (i.e. a set of ordered n-tuples of natural numbers), then the *characteristic function* of R is the n-place function f_R such that for any p,

$$f_R(\mathsf{p}) = \begin{cases} 1 & \text{if } R\mathsf{p} \text{ (i.e. if } \mathsf{p} \text{ is in } R), \\ 0 & \text{if not } R\mathsf{p}. \end{cases}$$

$f_=$ is thus the characteristic function of the identity relation. For any i, j, $f_=(i,j) = 1$ if $i = j$ and $f_=(i,j) = 0$ if $i \neq j$.

We shall call a function *Recursive* (capital 'R') if it can be obtained from the functions $+$, $\cdot$, $f_=$, and the various id^m_n by means of a finite number of applications of the two operations of composition and minimization of regular functions.

All Recursive functions are recursive, for $+$, $\cdot$, $f_=$ and the functions id^m_n are recursive, and the recursive functions are closed under composition and minimization of regular functions. On the other hand the zero function z is Recursive, for, as we saw in Chapter 7, z can be obtained from $\cdot$ by minimization. And s is obtainable by composition from Recursive functions and thus is Recursive too: for all i,

$$s(i) = i+1 = \mathrm{id}^1_1(i)+f_=(\mathrm{id}^1_1(i), \mathrm{id}^1_1(i)).$$

In the rest of Part II we show that all other recursive functions are also Recursive, and for this it suffices to show that if f and g are Recursive functions from which h is obtained by primitive recursion, then h is also Recursive.

We must first see that certain relations and functions are Recursive. A relation is Recursive iff its characteristic function is Recursive. (So $=$ is Recursive.)

Suppose, for example, that d is a two-place Recursive function and e is a three-place Recursive function. Let R be the 6-place relation defined by: Ri, j, k, m, n, q iff $d(j, n) = e(n, k, m)$. Then R is Recursive, for

$$f_R(i,j,k,m,n,q) = f_=(d(\mathrm{id}^6_2(i,j,k,m,n,q), \mathrm{id}^6_5(i,j,k,m,n,q)),$$
$$e(\mathrm{id}^6_5(i,j,k,m,n,q), \mathrm{id}^6_3(i,j,k,m,n,q), \mathrm{id}^6_4(i,j,k,m,n,q))).$$

Similarly, all other relations obtained by 'setting Recursive functions equal to each other' are Recursive.

Suppose that R and S are n-place Recursive relations. Then the intersection $(R \& S)$ of R and S and the complement $-R$ of R are Recursive, for

$$f_{(R\&S)}(\mathsf{p}) = f_R(\mathsf{p}) \cdot f_S(\mathsf{p}), \text{ and } f_{-R}(\mathsf{p}) = f_=(f_R(\mathsf{p}), z(f_R(\mathsf{p})))\ (= f_=(f_R(\mathsf{p}), 0)).$$

As & and $-$ suffice to define all truth-functional connectives, any relation obtained from Recursive relations by truth-functional, i.e., Boolean, operations is also Recursive. E.g. if Ri, j, k if and only if either $i = k$ or $k \neq j$, then R is Recursive.

If R is an $(n+1)$-place relation, then e will be said to be obtained from R by *minimization* if for any p, $e(\mathsf{p}) = \mu i R\mathsf{p}, i$. ($e$ may be undefined for some p.) An $(n+1)$-place relation will be called *regular* if for any p, there is an i such that $R\mathsf{p}, i$. The function obtained from a regular relation by minimization is everywhere defined.

If R is a regular, Recursive $(n+1)$-place relation, and e is obtained from R by minimization, then e is Recursive, for $e(\mathsf{p}) = \mu i f_{-R}(\mathsf{p}, i) = 0$. ($R\mathsf{p}, i$ iff $f_{-R}(\mathsf{p}, i) = 0$.)

Finally, if R is an $(n+1)$-place relation, then the $(n+1)$-place relation S will be said to be obtained from R by *bounded universal quantification*† if (for all p, j) $S\mathsf{p}, j$ iff $\forall i < j\, R\mathsf{p}, i$. If R is Recursive and S is obtained from R by bounded universal quantification, then S is Recursive. *Proof.* Let T be defined by: $T\mathsf{p}, j, i$ iff either not $R\mathsf{p}, i$ or $i = j$. T is regular (for all p, j, $T\mathsf{p}, j, j$) and Recursive (by the foregoing). Let d be defined by: $d(\mathsf{p}, j) = \mu i T\mathsf{p}, j, i$. d is Recursive. For any p, j, $d(\mathsf{p}, j) \leqslant j$. And $d(\mathsf{p}, j) = j$ iff for every $i < j$, $R\mathsf{p}, i$; iff $S\mathsf{p}, j$. So if e is defined by: $e(\mathsf{p}, j) = f_=(j, d(\mathsf{p}, j))$, then e is Recursive and the characteristic function of S.

S is said to be obtained from R by *bounded existential quantification*† if (for all p, j) $S\mathsf{p}, j$ iff $\exists i < j\, R\mathsf{p}, i$. Analogously, any relation obtained from a Recursive relation by bounded existential quantification is Recursive.

We'll now define J, the pairing function.

Definition

$$J(a, b) = \tfrac{1}{2}(a+b)(a+b+1)+a.$$

Lemma 14.1

J is a one–one function whose domain is the set of all ordered pairs $\langle a, b\rangle$ of natural numbers and whose range is the set of all natural numbers.

Proof. There are $n+1$ pairs $\langle a, b\rangle$ such that $a+b = n$ (*viz.*, $\langle 0, n\rangle$, $\langle 1, n-1\rangle, \ldots, \langle n, 0\rangle$). So there are $0+1+2+\ldots+n$, $= \frac{1}{2}n(n+1)$, pairs $\langle c, d\rangle$ such that $c+d < n$. We'll say that $\langle c, d\rangle$ *precedes* $\langle a, b\rangle$ *in order O* (cf. Chapter 13) if either $c+d < a+b$ or ($c+d = a+b$ and $c < a$). There are a natural numbers less than a. So if $a+b = n$, there are $\frac{1}{2}n(n+1)+a$ pairs that precede $\langle a, b\rangle$ in order O. But if $a+b = n$, then

$$\tfrac{1}{2}n(n+1)+a = J(a, b).$$

So $J(a, b)$ is precisely the number of pairs preceding $\langle a, b\rangle$ in order O.

$a, b \leqslant J(a, b)$. J is Recursive, for J is obtained from a regular Recursive function by minimization: $J(a, b) = \mu i[i+i = (a+b)(a+b+1)+2a]$.

Define K and L, the inverse pairing functions, by:

$K(i) = \mu a \exists b \leqslant i J(a, b) = i$ (i.e. $\mu a[\exists b < i J(a, b) = i \vee J(a, i) = i]$),
$L(i) = \mu b \exists a \leqslant i J(a, b) = i$. By Lemma 14.1, K and L are Recursive.

† These definitions differ slightly from those given in Chapter 7 in that '$<$' is used instead of '$\leqslant$'.

We now define some more relations and functions; it should be evident from their definitions that they are Recursive.

m *divides* $n \leftrightarrow \exists i \leqslant n\ i \cdot m = n$.

p *is prime* $\leftrightarrow \{p \neq 0 \,\&\, p \neq 1 \,\&\, \forall m \leqslant p[m \text{ divides } p \rightarrow (m = 1 \vee m = p)]\}$.

$m < n \leftrightarrow \exists i < n\ i = m$.

$m \dot{-} n = \mu i([n < m \rightarrow n + i = m] \,\&\, [-n < m \rightarrow i = 0])$.

n *is a power of the prime* $p \leftrightarrow \{n \neq 0 \,\&\, p \text{ is prime} \,\&\, \forall m \leqslant n[m \text{ divides } n$
$\rightarrow (m = 1 \vee p \text{ divides } m)]\}$.

(Notice that we can't simply say that n is a power of k iff for every $m \leqslant n$, if m divides n, then $m = 1$ or k divides m; let $n = k = 6$, $m = 2$.)

$\eta(p, b) = \mu i[(p \text{ is prime} \,\&\, i \text{ is a power of the prime } p \,\&\, i > b \,\&\, i > 1)$
$\vee\ (p \text{ is not prime} \,\&\, i = 0)]$.

For prime p, $\eta(p, b)$ is the least number whose base p numeral is *longer* than the base p numeral for b. E.g. $\eta(7, 25) = 49$. (Note that $25 = 34_7$ and $49 = 100_7$.)

$a \underset{p}{*} b = a \cdot \eta(p, b) + b$.

If $a \neq 0$, $a \underset{p}{*} b$ is $\neq 0$ and is the number denoted in base p notation by the result of writing the base p numeral for b directly to the right of that for a. So, e.g. $4 \underset{7}{*} 25 = 4 \cdot 49 + 25 = 221$, and $4_7 = 4$, $34_7 = 25$, and $434_7 = 221$. In what follows, association is assumed to be to the left: '$a \underset{p}{*} b \underset{p}{*} c$' means '$(a \underset{p}{*} b) \underset{p}{*} c$', not '$a \underset{p}{*} (b \underset{p}{*} c)$'. Then if $a \neq 0$, $a \underset{p}{*} b \underset{p}{*} c \underset{p}{*} \ldots \underset{p}{*} z$ is the number denoted in base p notation by the result of writing down the base p numeral for b directly to the right of that for a, then that for c directly to the right of *that*, ... and then that for z directly to the right of *that*.

$a\ \mathit{part}_p\ b \leftrightarrow \exists c \leqslant b\, \exists d \leqslant b\, [c \underset{p}{*} a \underset{p}{*} d = b \vee c \underset{p}{*} a = b \vee a \underset{p}{*} d = b \vee a = b]$.

a part$_p$ b iff $a = 0$ or $a = b$ or b's base p numeral can be obtained by attaching base p numerals to the left and/or right of a's base p numeral.

$\alpha(p, q, j) = \mu i[(p \dot{-} 1) \underset{p}{*} j \underset{p}{*} i \text{ part}_p\, q \vee i = q]$. ('$i = q$' is for 'waste cases'.)

$\beta(i, j) = \alpha(K(i), L(i), j)$.

Lemma 14.2. (*The β-function lemma*)
For any k and any finite sequence of natural numbers $i_0, \ldots, i_k$, there exists a natural number i such that for every $j \leqslant k$, $\beta(i, j) = i_j$.

Proof. Let $i_0, \ldots, i_k$ be a finite sequence of natural numbers. Let p be a prime such that $p - 1$ is greater than all of $i_0, \ldots, i_k, k$. (There are infinitely many primes.) Let $s = p - 1$. $s \neq 0$. All of $s, 0, i_0, \ldots, k, i_k$ are represented by single digits in base p notation (!). Let

$$q = s \underset{p}{*} 0 \underset{p}{*} i_0 \underset{p}{*} s \underset{p}{*} 1 \underset{p}{*} i_1 \underset{p}{*} \ldots \underset{p}{*} s \underset{p}{*} k \underset{p}{*} i_k.$$

Then for every $j \leqslant k$, $\alpha(p, q, j) = i_j$. Let $i = J(p, q)$. Then for every $j \leqslant k$, $\beta(i, j) = i_j$.

Suppose now that f is an n-place function, that g is an $(n+2)$-place function, and that h is obtained from f and g by primitive recursion. Then $h(\mathsf{p}, 0) = f(\mathsf{p})$ and (for any k) for every $j < k$, $h(\mathsf{p}, j') = g(\mathsf{p}, j, h(\mathsf{p}, j))$. By the β-function lemma, for any k there is an i such that for every $j \leqslant k$, $\beta(i, j) = h(\mathsf{p}, j)$. These is are precisely those such that $\beta(i, 0) = f(\mathsf{p})$ and for every $j < k$, $\beta(i, j') = g(\mathsf{p}, j, \beta(i, j))$. Therefore, if R is the $(n+2)$-place relation defined by:

$$R\mathsf{p}, k, i \text{ iff } \beta(i, 0) = f(\mathsf{p}) \,\&\, \forall j < k\ \beta(i, j') = g(\mathsf{p}, j, \beta(i, j)),$$

then R is regular; and R is Recursive if f and g are. So if d is the $(n+1)$-place function defined by: $d(\mathsf{p}, k) = \mu i R\mathsf{p}, k, i$, then d is Recursive if f and g are. Moreover $d(\mathsf{p}, k)$ is the least i such that for every $j \leqslant k$, $\beta(i, j) = h(\mathsf{p}, j)$. For any such i, $\beta(i, k) = h(\mathsf{p}, k)$. We may thus define h by composition from β, d, id^{n+1}_{n+1}: $h(\mathsf{p}, k) = \beta(d(\mathsf{p}, k), \mathrm{id}^{n+1}_{n+1}(\mathsf{p}, k))$. As β and id^{n+1}_{n+1} are Recursive, h is Recursive if f and g are Recursive.

Thus any function obtained by primitive recursion from Recursive functions is itself Recursive.

We have therefore shown that a function is recursive if and only if it is Recursive.

Part III

We'll now show that all Recursive functions are representable in Q, from which we conclude that all recursive functions are representable in Q.

The identity functions id^m_n are all representable in Q: since for any $i_1, \ldots, i_m$, $\forall x_{m+1}((\mathbf{i_1} = \mathbf{i_1} \,\&\, \ldots \,\&\, \mathbf{i_m} = \mathbf{i_m} \,\&\, x_{m+1} = \mathbf{i_n}) \leftrightarrow x_{m+1} = \mathbf{i_n})$ is *valid*,

$$(x_1 = x_1 \,\&\, \ldots \,\&\, x_m = x_m \,\&\, x_{m+1} = x_n)$$

represents id^m_n in Q.

We now show that addition is represented in Q by the formula

$$x_1 + x_2 = x_3.$$

Lemma 14.3

Suppose that $i+j = k$. Then $\vdash_Q \mathbf{i}+\mathbf{j} = \mathbf{k}$.

Proof. The proof is an induction on j. Basis step: $j = 0$. We must show that $\vdash_Q \mathbf{i}+\mathbf{0} = \mathbf{i}$. But this follows from $Q4$. Induction step: $j = m'$. Then for some n, $k = n'$ and $i+m = n$, whence by the induction hypothesis, $\vdash_Q \mathbf{i}+\mathbf{m} = \mathbf{n}$, and therefore $\vdash_Q (\mathbf{i}+\mathbf{m})' = \mathbf{n}'$. Since

$$\vdash_Q (\mathbf{i}+\mathbf{m})' = \mathbf{i}+\mathbf{m}'$$

by $Q5$, it follows that $\vdash_Q \mathbf{i}+\mathbf{j} = \mathbf{k}$.

Lemma 14.4

$x_1 + x_2 = x_3$ represents addition in Q.

Proof. $\forall x_3(\mathbf{i}+\mathbf{j} = x_3 \leftrightarrow x_3 = \mathbf{k})$ is a logical consequence of $\mathbf{i}+\mathbf{j} = \mathbf{k}$, which, by 14.3, is a theorem of Q if $i+j = k$.

Multiplication:

Lemma 14.5

Suppose that $i \cdot j = k$. Then $\vdash_Q \mathbf{i} \cdot \mathbf{j} = \mathbf{k}$.

Proof. Induction on j. If $j = 0$, we must show that $\vdash_Q \mathbf{i} \cdot \mathbf{0} = \mathbf{0}$. But this follows from $Q6$. If $j = m'$, then $k = n+i$, where $n = i \cdot m$. By the hypothesis of the induction, $\vdash_Q \mathbf{i} \cdot \mathbf{m} = \mathbf{n}$. By 14.3, $\vdash_Q \mathbf{n}+\mathbf{i} = \mathbf{k}$. By $Q7$, $\vdash_Q \mathbf{i} \cdot \mathbf{m}' = \mathbf{i} \cdot \mathbf{m}+\mathbf{i}$. So $\vdash_Q \mathbf{i} \cdot \mathbf{m}' = \mathbf{k}$, i.e., $\vdash_Q \mathbf{i} \cdot \mathbf{j} = \mathbf{k}$.

Lemma 14.6

$x_1 \cdot x_2 = x_3$ represents multiplication in Q.

Proof. This follows from 14.5 just as 14.4 followed from 14.3.

So $+$ and $\cdot$ are representable in Q.

Let's now verify that if $i \neq j$, then $\mathbf{i} \neq \mathbf{j}$ is a theorem of Q.

Lemma 14.7

If $i \neq j$, then $\vdash_Q \mathbf{i} \neq \mathbf{j}$.

Proof. We may suppose without loss of generality that $i < j$. Induction on i. If $i = 0$, then $j > 0$, and so for some n, $j = n'$. We must show that $\vdash_Q \mathbf{0} \neq \mathbf{j}$, i.e., that $\vdash_Q \mathbf{0} \neq \mathbf{n}'$. But this immediately follows from $Q2$. If $i = m'$, then $j = n'$ and $m < n$, for some n. By the induction hypothesis, $\vdash_Q \mathbf{m} \neq \mathbf{n}$, and hence by $Q1$, $\vdash_Q \mathbf{m}' \neq \mathbf{n}'$, i.e., $\vdash_Q \mathbf{i} \neq \mathbf{j}$.

Lemma 14.8

Let $A(x_1, x_2, x_3) =$ the formula

$$(x_1 = x_2 \,\&\, x_3 = \mathbf{1}) \vee (x_1 \neq x_2 \,\&\, x_3 = \mathbf{0}).$$

Then $A(x_1, x_2, x_3)$ represents $f_=$ in Q.

Proof. If $f_=(i,j) = 1$, then $i = j$. So $\vdash_Q \mathbf{i} = \mathbf{j} \,\&\, \mathbf{1} = \mathbf{1}$, so $\vdash_Q A(\mathbf{i}, \mathbf{j}, \mathbf{1})$, whence $\vdash_Q \forall x_3(A(\mathbf{i}, \mathbf{j}, x_3) \leftrightarrow x_3 = \mathbf{1})$, as $\forall x_3(A(\mathbf{i}, \mathbf{j}, x_3) \rightarrow x_3 = \mathbf{1})$ is a logical consequence of $A(\mathbf{i}, \mathbf{j}, \mathbf{1})$ when $i = j$. If $f_=(i,j) = 0$, then $i \neq j$. By 14.7 $\vdash_Q \mathbf{i} \neq \mathbf{j}$. So $\vdash_Q \mathbf{i} \neq \mathbf{j} \,\&\, \mathbf{0} = \mathbf{0}$, whence $\vdash_Q \forall x_3(A(\mathbf{i}, \mathbf{j}, x_3) \leftrightarrow x_3 = \mathbf{0})$.

Thus $f_=$ is also representable in Q. We now show that any function obtained by composition from functions representable in Q is also representable in Q.

Suppose that $A(x_1, \ldots, x_m, x)$ represents f in Q, and that

$$B_1(\mathsf{x}, x_{n+1}), \ldots, B_m(\mathsf{x}, x_{n+1})$$

represent $g_1, \ldots, g_m$, respectively. Then if h is obtained from $f, g_1, \ldots, g_m$, by composition,

$$C(\mathsf{x}, x), = \exists y_1 \ldots \exists y_m(B_1(\mathsf{x}, y_1) \,\&\, \ldots \,\&\, B_m(\mathsf{x}, y_m) \,\&\, A(y_1, \ldots, y_m, x)),$$

reprêsents h.

For if $g_1(\mathsf{p}) = i_1, \ldots, g_m(\mathsf{p}) = i_m$, and $f(i_1, \ldots, i_m) = j$, then $h(\mathsf{p}) = j$, and

$$\vdash_Q B_1(\mathbf{p}, \mathbf{i}_1), \tag{1}$$

$$\vdash_Q \forall x_{n+1}(B_1(\mathbf{p}, x_{n+1}) \rightarrow x_{n+1} = \mathbf{i}_1), \tag{2}$$

$$\vdots$$

$$\vdash_Q B_m(\mathbf{p}, \mathbf{i}_\mathbf{m}), \tag{2m−1}$$

$$\vdash_Q \forall x_{n+1}(B_m(\mathbf{p}, x_{n+1}) \rightarrow x_{n+1} = \mathbf{i}_\mathbf{m}), \tag{2m}$$

$$\vdash_Q A(\mathbf{i}_1, \ldots, \mathbf{i}_\mathbf{m}, \mathbf{j}), \text{ and} \tag{2m+1}$$

$$\vdash_Q \forall x(A(\mathbf{i}_1, \ldots, \mathbf{i}_\mathbf{m}, x) \rightarrow x = \mathbf{j}). \tag{2m+2}$$

(1), (3), ..., $(2m-1)$, and $(2m+1)$ clearly entail that $\vdash_Q C(\mathbf{p}, \mathbf{j})$. And (2), (4), ..., $(2m)$, and $(2m+2)$ entail that $\vdash_Q \forall x(C(\mathbf{p}, x) \to x = \mathbf{j})$. We may see this as follows: Assume we have $B_1(\mathbf{p}, y_1), \ldots, B_m(\mathbf{p}, y_m)$, and $A(y_1, \ldots, y_m, x)$. From (2), we have $y_1 = \mathbf{i_1}, \ldots,$ and from $(2m)$ we have $y_m = \mathbf{i_m}$. So we have $A(\mathbf{i_1}, \ldots, \mathbf{i_m}, x)$, whence from $(2m+2)$ we have $x = \mathbf{j}$. Thus

$$\vdash_Q \forall x(\exists y_1, \ldots \exists y_m(B_1(\mathbf{p}, y_1) \,\&\, \ldots \,\&\, B_m(\mathbf{p}, y_m) \,\&\, A(y_1, \ldots, y_m, x)) \to x = \mathbf{j}),$$

i.e. $\vdash_Q \forall x(C(\mathbf{p}, x) \to x = \mathbf{j})$, and therefore C represents h.

Lemma 14.9

For each i, $\vdash_Q \forall x\, x' + \mathbf{i} = x + \mathbf{i}'$.

Proof. Induction on i. If $i = 0$, $\forall x\, x' + \mathbf{o} = x + \mathbf{o}'$ follows from

$$\forall x(x' + \mathbf{o} = x' = (x + 0)' = x + \mathbf{o}'),$$

which follows from $Q4$ and $Q5$. If $i = m'$, then by the induction hypothesis $\vdash_Q \forall x\, x' + \mathbf{m} = x + \mathbf{m}'$, whence by $Q5$, $\vdash_Q \forall x(x' + \mathbf{m}' = (x' + \mathbf{m})' = (x + \mathbf{m}')' = x + \mathbf{m}'')$, and hence

$$\vdash_Q \forall x\, x' + \mathbf{i} = x + \mathbf{i}'.$$

We now define $x_1 < x_2$ to be the formula $\exists x_3\, x_3' + x_1 = x_2$.

Lemma 14.10

If $i < j$, then $\vdash_Q \mathbf{i} < \mathbf{j}$.

Proof. Suppose $i < j$. Then for some m, $m' + i = j$. By 14.3,

$$\vdash_Q \mathbf{m}' + \mathbf{i} = \mathbf{j}, \quad \text{and so} \quad \vdash_Q \exists x_3 x_3' + \mathbf{i} = \mathbf{j}, \text{i.e.} \vdash_Q \mathbf{i} < \mathbf{j}.$$

Lemma 14.11

For each i, $\vdash_Q \forall x(x < \mathbf{i} \to x = \mathbf{o} \vee \ldots \vee x = \mathbf{i} - \mathbf{1})$ (where, if $i = 0$, the consequent is an empty disjunction and hence is to be regarded as equivalent to $\mathbf{o} \neq \mathbf{o}$).

Proof. Induction on i. Basis step: $i = 0$. We must show $\vdash_Q \forall x - x < \mathbf{o}$. By $Q3$ we have $x = \mathbf{o} \vee \exists y\, x = y'$. Assume $x < \mathbf{o}$, i.e., $\exists w\, w' + x = \mathbf{o}$. If $x = \mathbf{o}$ holds, we have $w' = w' + \mathbf{o}$ (by $Q4$) $= w' + x = \mathbf{o}$, which is impossible by $Q2$. If $x = y'$ holds, we have $(w' + y)' = w' + y'$ (by $Q5$) $= w' + x = \mathbf{o}$ which is again impossible by $Q2$. Thus $\vdash_Q \forall x - x < \mathbf{o}$.

Induction step. We suppose $\vdash_Q \forall x(x < \mathbf{i} \to x = \mathbf{o} \vee \ldots \vee x = \mathbf{i} - \mathbf{1})$. We must show $\vdash_Q \forall x(x < \mathbf{i}' \to x = \mathbf{o} \vee x = \mathbf{o}' \vee \ldots \vee x = \mathbf{i})$. Assume we have

$x < \mathbf{i}'$, i.e., $\exists w\, w' + x = \mathbf{i}'$. By $Q3$ we have $\exists y\, x = y' \vee x = \mathbf{o}$. If $x = y'$ holds, then we have $\mathbf{i}' = w' + x = w' + y' = (w' + y)'$ (by $Q5$), whence by $Q1$ we have $\mathbf{i} = w' + y$, and therefore $y < \mathbf{i}$. By the induction hypothesis we have $y = \mathbf{o} \vee \ldots \vee y = \mathbf{i} - \mathbf{1}$ (if $i = 0$, we have $\mathbf{o} \neq \mathbf{o}$), and therefore we have $x = \mathbf{o}' \vee \ldots \vee x = \mathbf{i}$ (if $i = 0$, we have $\mathbf{o} \neq \mathbf{o}$), and therefore we have $x = \mathbf{o} \vee x = \mathbf{o}' \vee \ldots \vee x = \mathbf{i}$, which we also have in case $x = \mathbf{o}$ holds.

Lemma 14.12

For each i, $\vdash_Q \forall x(\mathbf{i} < x \to x = \mathbf{i}' \vee \mathbf{i}' < x)$.

Proof. Assume $\mathbf{i} < x$, i.e. $\exists w\, w' + \mathbf{i} = x$. We have $w = \mathbf{o} \vee \exists y\, w = y'$ by $Q3$. From $w = \mathbf{o}$ and $w' + \mathbf{i} = x$, we have $\mathbf{o}' + \mathbf{i} = x$, whence by 14.3 we have $x = \mathbf{i}'$. From $w = y'$ and $w' + \mathbf{i} = x$, we have $y'' + \mathbf{i} = x$, whence by 14.9 we have $y' + \mathbf{i}' = x$, and so we have $\mathbf{i}' < x$.

Lemma 14.13

For each i, $\vdash_Q \forall x(\mathbf{i} < x \vee x = \mathbf{i} \vee x < \mathbf{i})$.

Proof. Induction on i. Basis step: $i = 0$. Assume $x \neq \mathbf{o}$. By $Q3$ we then have $\exists y\, x = y'$, and so by $Q4$ we have $\exists y\, y' + \mathbf{o} = x$, i.e., $\mathbf{o} < x$. Induction step: we suppose $\vdash_Q \forall x(\mathbf{i} < x \vee x = \mathbf{i} \vee x < \mathbf{i})$. We must show $\vdash_Q \forall x(\mathbf{i}' < x \vee x = \mathbf{i}' \vee x < \mathbf{i}')$. By 14.12

$$\vdash_Q \forall x(\mathbf{i} < x \to x = \mathbf{i}' \vee \mathbf{i}' < x). \text{ By 14.10} \tag{I}$$

$$\vdash_Q \forall x(x = \mathbf{i} \to x < \mathbf{i}'). \text{ And by 14.11 and 14.10} \tag{II}$$

$$\vdash_Q \forall x(x < \mathbf{i} \to x < \mathbf{i}'). \tag{III}$$

But from (I), (II), (III), and the induction hypothesis, it follows that

$$\vdash_Q \forall x(\mathbf{i}' < x \vee x = \mathbf{i}' \vee x < \mathbf{i}').$$

We can now show that the result g of applying minimization to any regular $(n+1)$-place function f that is representable in Q is also representable in Q.

Suppose that f is a regular $(n+1)$-place function, and that

$$A(\mathsf{x}, x_{n+1}, x_{n+2})$$

represents f in Q. Let $B(\mathsf{x}, x_{n+1})$ = the formula

$$(A(\mathsf{x}, x_{n+1}, \mathbf{o}) \,\&\, \forall w(w < x_{n+1} \to -A(\mathsf{x}, w, \mathbf{o}))).$$

Then B represents g in Q.

For suppose that $g(p) = i$. Then $f(p, i) = 0$, and for any $j < i$, $f(p, j) \neq 0$. Since A represents f in Q, we have

$$\vdash_Q A(\mathbf{p}, \mathbf{i}, \mathbf{0}), \text{ and (if } i > 0), \qquad (i)$$

$$\vdash_Q -A(\mathbf{p}, \mathbf{0}, \mathbf{0}), \qquad (0)$$

$$\vdots \qquad\qquad \vdots$$

$$\vdash_Q -A(\mathbf{p}, \mathbf{i-1}, \mathbf{0}). \qquad (i-1)$$

$(0), \ldots, (i-1)$, and 14.11 entail that

$$\vdash_Q \forall w(w < \mathbf{i} \rightarrow -A(\mathbf{p}, w, \mathbf{0})), \qquad (i+1)$$

which, together with (i), entails that $\vdash_Q B(\mathbf{p}, \mathbf{i})$.

We must show that $\vdash_Q \forall x_{n+1}(B(\mathbf{p}, x_{n+1}) \rightarrow x_{n+1} = \mathbf{i})$. Assume $B(\mathbf{p}, x_{n+1})$, i.e., $A(\mathbf{p}, x_{n+1}, \mathbf{0})$ & $\forall w(w < x_{n+1} \rightarrow -A(\mathbf{p}, w, \mathbf{0}))$. From (i) and

$$\forall w(w < x_{n+1} \rightarrow -A(\mathbf{p}, w, \mathbf{0})), \text{ we have } -\mathbf{i} < x_{n+1}.$$

From $A(\mathbf{p}, x_{n+1}, \mathbf{0})$ and $(i+1)$, we have $-x_{n+1} < \mathbf{i}$. Thus by 14.13 we have $x_{n-1} = \mathbf{i}$. So $\vdash_Q \forall x_{n+1}(B(\mathbf{p}, x_{n+1}) \rightarrow x_{n+1} = \mathbf{i})$.

Exercises

14.1 Verify the following assertion: all recursive functions are representable in the theory ('R') whose language is L and whose theorems are the consequences in L of the following infinitely many sentences:

$\mathbf{i} \neq \mathbf{j}$ for all i, j such that $i \neq j$;

$\mathbf{i} + \mathbf{j} = \mathbf{k}$ for all i, j, k such that $i + j = k$;

$\mathbf{i} \cdot \mathbf{j} = \mathbf{k}$ for all i, j, k such that $i \cdot j = k$;

$\forall x(x < \mathbf{i} \rightarrow x = \mathbf{0} \vee \ldots \vee x = \mathbf{i-1})$ for all i;

and $\forall x(x < \mathbf{i} \vee x = \mathbf{i} \vee \mathbf{i} < x)$, for all i.

14.2 Show that none of the following sentences are theorems of Q:

(*a*) $\forall x\, x \neq x'$,

(*b*) $\forall x \forall y \forall z\, x + (y + z) = (x + y) + z$,

(*c*) $\forall x \forall y\, x + y = y + x$,

(*d*) $\forall x\, \mathbf{0} + x = x$,

(*e*) $\forall x\, x < x'$,

(*f*) $\forall x \forall y -(x < y \,\&\, y < x)$,

(*g*) $\forall x \forall y \forall z\, x \cdot (y \cdot z) = (x \cdot y) \cdot z$,

(*h*) $\forall x \forall y\, x \cdot y = y \cdot x$,

(*i*) $\forall x\, \mathbf{0} \cdot x = 0$,

(*j*) $\forall x \forall y \forall z\, x \cdot (y + z) = x \cdot y + x \cdot z$.

Hint: Let a and b be two objects that are not natural numbers, and consider the following successor, addition, and multiplication tables:

x	x'
i	i'
a	a
b	b

$\curvearrowright +$	j	a	b
i	$i+j$	b	a
a	a	b	a
b	b	b	a

$\curvearrowright \cdot$	0	$j \neq 0$	a	b
0	0	0	a	b
$i \neq 0$	0	$i \cdot j$	a	b
a	0	b	b	b
b	0	a	a	a

15
Undecidability, indefinability and incompleteness

We are now in a position to give a unified treatment of some of the central negative results of logic: Church's theorem on the undecidability of logic, Tarski's theorem on the indefinability of truth, and Gödel's first theorem on the incompleteness of systems of arithmetic. These theorems can all be seen as more or less direct consequences of the result of the last chapter, that all recursive functions are representable in Q, and a certain exceedingly ingenious lemma ('the diagonal lemma'), the idea of which is due to Gödel, and which we shall prove below. The first notion that we have to introduce is that of a *gödel numbering.*

A *gödel numbering* is an assignment of natural numbers (called 'gödel numbers') to expressions (in some set) that meets these conditions: (1) different gödel numbers are assigned to different expressions: (2) it is effectively calculable what the gödel number of any expression is; (3) it is effectively decidable whether a number is the gödel number of some expression in the set, and, if so, effectively calculable which expression it is the gödel number of.

Gödel numberings enable one to regard interpreted languages supposed to be 'about' the natural numbers – i.e. having the set of natural numbers as the domain of their intended interpretation – as also referring to the numbered expressions. The possibility then arises that certain sentences, ostensibly referring to certain numbers, could be seen as referring, via the gödel numbering, to certain expressions that are *identical* with those very sentences themselves. The state of affairs just described is no mere possibility; the proof of the diagonal lemma shows how it arises, and succeeding theorems show how it may be exploited.

We shall consider a particular set of expressions and a particular gödel numbering, to which we appropriate the words 'expression' and 'gödel number'. There is nothing special about our particular gödel numbering; the theorems and proofs that we are going to give with respect to the one we use could have been given with respect to any number of others. Our expressions are finite sequences of these (distinct) symbols.

We'll make the following 'conventions' about the identity of certain symbols: we stipulate that $x_0 = x$, $x_1 = y$, $f_0^0 = \mathbf{o}$, $f_0^1 = {}'$, $f_0^2 = +$, $f_1^2 = \cdot$,

TABLE 15-1

(	)	&	∃	x_0	f_0^0	f_0^1	f_0^2	…	A_0^0	A_0^1	A_0^2	…
	,	v	∀	x_1	f_1^0	f_1^1	f_1^2	…	A_1^0	A_1^1	A_1^2	…
		—		x_2	f_2^0	f_2^1	f_2^2	…	A_2^0	A_2^1	A_2^2	…
		↔		.	.	.	.		.	.	.	
		→		.	.	.	.		.	.	.	
				.	.	.	.		.	.	.	

and $A_0^2 = \;=$. We now assign each symbol in Table 15-1 the number in the corresponding location in Table 15-2 as its gödel number:

TABLE 15-2

1	2	3	4	5	6	68	688	…	7	78	788	…
	29	39	49	59	69	689	6889	…	79	789	7889	…
		399		599	699	6899	68899	…	799	7899	78899	…
		3999		.	.	.	.		.	.	.	
		39999		.	.	.	.		.	.	.	
				.	.	.	.		.	.	.	

We'll write 'gn' to mean 'the gödel number of'. Thus,

$$\text{gn}(x) = 5, \text{gn}(y) = 59, \text{gn}(\mathbf{o}) = 6, \text{gn}(') = 68, \text{gn}(+) = 688,$$
$$\text{gn}(\cdot) = 6889, \quad \text{and} \quad \text{gn}(=) = 788.$$

We must now extend the gödel numbering so that all finite sequences of symbols in Table 15-1 are assigned gödel numbers. (We don't distinguish between a single symbol and the sequence which consists of that one symbol.) The principle can be indicated in a single example: Since gn(∃) = 4, gn(x) = 5, gn(( ) = 1, and gn(=) = 788, we want

$$\text{gn}(\exists x(x =)$$

to be 4515788.

The principle is that if expression A has gödel number i, and B has j, then AB, the expression formed by writing A immediately before B, is to have as its gödel number the number denoted by the decimal arabic numeral formed by writing the decimal arabic numeral for i immediately before the decimal arabic numeral for j. It's clear that our gödel numbering really is a gödel numbering in the sense of the second paragraph.

We now introduce the notion of the *diagonalization* of an expression A. Recall from the last chapter that $\mathbf{n}$ is the result of writing n occurrences of $'$ immediately after $\mathbf{o}$. If A is an expression with gödel number n, we define $\ulcorner A \urcorner$ to be the expression $\mathbf{n}$. In what follows $\ulcorner A \urcorner$ will be seen to behave rather like a name for the expression A. The *diagonalization* of A is the expression

$$\exists x (x = \ulcorner A \urcorner \,\&\, A).$$

If A is a formula in the language of arithmetic that contains just the variable x free, then the diagonalization of A will be a sentence that 'says that' A is true of its own gödel number – or, more precisely, the diagonalization will be true (in the standard interpretation $\mathcal{N}$) if and only if A is true (in $\mathcal{N}$) of its own gödel number.

Lemma 1

There is a recursive function, diag, such that if n is the gödel number of an expression A, diag (n) is the gödel number of the diagonalization of A.

Proof. Let lh ('length') be defined by lh $(n) = \mu m(0 < m \,\&\, n < 10^m)$. Since every natural number n is less than 10^m for some positive m, and as exponentiation and less than are recursive, lh is a recursive function. lh (n) is the number of digits in the usual arabic numeral for n. Thus lh $(1879) = 4$; lh $(0) =$ lh $(9) = 1$.

Let $*$ be defined by: $m * n = m \cdot 10^{\mathrm{lh}(n)} + n$. $*$ is recursive. If $m \neq 0$, $m * n$ is the number denoted by the arabic numeral formed by writing the arabic numeral for m immediately before the arabic numeral for n. Thus $2 * 3 = 23$.

Let num be defined by: num $(0) = 6$; num $(n+1) =$ num $(n) * 68$ (all n). num is recursive. As gn $(\mathbf{o}) = 6$ and gn $(') = 68$, num (n) is the gödel number of $\mathbf{n}$.

As no arabic numeral for a gödel number contains the digit '0', diag (n) may be taken to $= 4515788 * (\mathrm{num}(n) * (3 * (n * 2)))$. diag is then recursive.

We'll now reserve the word 'theory' for those theories whose variables are $x_0, x_1, \ldots,$ and whose names, n-place function signs, sentence letters, and n-place predicate letters are some (or all) of $f_0^0, f_1^0, \ldots, f_0^n, f_1^n, \ldots,$ $A_0^0, A_0^1, \ldots, A_0^n, A_1^n, \ldots,$ respectively. As in the last chapter, we assume that $\mathbf{o}$ and $'$ occur in the language of all theories. We assume further that the set of gödel numbers of symbols of the language of the theory is recursive, i.e. (by Church's thesis), that there is an effective procedure for deciding whether a given symbol may occur in some sentence in the language of the theory.

Here's the diagonal lemma:

Lemma 2

Let T be a theory in which diag is representable. Then for any formula $B(y)$ (of the language of T, containing just the variable y free), there is a sentence G such that

$$\vdash_T G \leftrightarrow B(\ulcorner G \urcorner).$$

Proof. Let $A(x, y)$ represent diag in T. Then for any n, k, if diag $(n) = k$, $\vdash_T \forall y (A(\mathbf{n}, y) \leftrightarrow y = \mathbf{k})$.

Let F be the expression $\exists y (A(x, y) \,\&\, B(y))$. F is a formula of the language of T that contains just the variable x free.

Let n be the gödel number of F.

Let G be the expression $\exists x (x = \mathbf{n} \,\&\, \exists y (A(x, y) \,\&\, B(y)))$. As $\mathbf{n} = \ulcorner F \urcorner$, G is the diagonalization of F and a sentence of the language of T. Since G is logically equivalent to $\exists y (A(\mathbf{n}, y) \,\&\, B(y))$, we have

$$\vdash_T G \leftrightarrow \exists y (A(\mathbf{n}, y) \,\&\, B(y)).$$

Let k be the gödel number of G. Then

$$\mathrm{diag}(n) = k, \quad \text{and} \quad \mathbf{k} = \ulcorner G \urcorner.$$

So $\quad \vdash_T \forall y (A(\mathbf{n}, y) \leftrightarrow y = \mathbf{k})$.

So $\quad \vdash_T G \leftrightarrow \exists y (y = \mathbf{k} \,\&\, B(y))$.

So $\quad \vdash_T G \leftrightarrow B(\mathbf{k})$, i.e., $\vdash_T G \leftrightarrow B(\ulcorner G \urcorner)$.

A theory is called *consistent* if there is no theorem of the theory whose negation is also a theorem. Equivalently, a theory is consistent iff there is some sentence in its language that is not a theorem, iff the theory is satisfiable.

A set θ of natural numbers is said to be *definable in* theory T if there is a formula $B(x)$ of the language of T such that for any number k, if $k \in \theta$, then $\vdash_T B(\mathbf{k})$, and if $k \notin \theta$, then $\vdash_T -B(\mathbf{k})$. The formula $B(x)$ is said to define θ in T. A two-place relation R of natural numbers is likewise definable in T if there is a formula $C(x, y)$ of the language of T such that for any numbers k, n, if kRn, then $\vdash_T C(\mathbf{k}, \mathbf{n})$, and if $k \not R n$, then $\vdash_T -C(\mathbf{k}, \mathbf{n})$, and $C(x, y)$ is then said to define R in T. (A perfectly analogous definition of definability can be given for three- and more-place relations on natural numbers; we won't need this more general notion, however.)

A theory T is called an *extension* of theory S if S is a subset of T, i.e., if any theorem of S is a theorem of T. If f is a function that is representable in S, and T is an extension of S, then f is representable in T, and indeed is represented in T by the same formula that represents it in S. Similarly, any formula that defines a set in some theory defines it in any extension of that theory.

Lemma 3

If T is a consistent extension of Q, then the set of gödel numbers of theorems of T is not definable in T.

Proof. Let T be an extension of Q. Then diag is representable in T; for as diag is a recursive function, and all recursive functions are representable in Q, diag is representable in Q, and hence is representable in any extension of Q.

Suppose now that $C(y)$ defines the set θ of gödel numbers of theorems of T. By the diagonal lemma, there is a sentence G such that

$$\vdash_T G \leftrightarrow -C(\ulcorner G \urcorner).$$

Let $k = \text{gn}(G)$. Then

$$\vdash_T G \leftrightarrow -C(\mathbf{k}). \qquad (*)$$

Then $\vdash_T G$. For if G is not a theorem of T, then $k \notin \theta$, and so, as $C(y)$ defines θ, $\vdash_T -C(\mathbf{k})$, whence by (*), $\vdash_T G$.

So $k \in \theta$. So $\vdash_T C(\mathbf{k})$, as $C(y)$ defines θ. So, by (*), $\vdash_T -G$, and T is therefore inconsistent.

A set of expressions is called *decidable* if the set of gödel numbers of its members is a recursive set. Thus a theory T is decidable iff the set θ of gödel numbers of its theorems is recursive, iff the characteristic function of θ is recursive.

If a theory is decidable, then an effective method exists for deciding whether any given sentence is a theorem of the theory. For to determine whether a sentence is a theorem, calculate its gödel number first and then calculate the value of the (recursive, hence calculable) characteristic function for the gödel number as argument. The sentence is a theorem iff the value is 1.

Conversely, if a theory is not decidable, then *unless Church's thesis is false*, no effective method exists for deciding whether a given sentence is a theorem of the theory. For if there were such a method, then the characteristic function of the set of gödel numbers of theorems would also be effectively calculable, and hence recursive, by Church's thesis.

Theorem 1

No consistent extension of Q is decidable.

Proof. Suppose T is a consistent extension of Q. Then by Lemma 3, the set θ of gödel numbers of theorems of T is not definable in T. Now if $A(x,y)$ represented the characteristic function f of θ in T, then $A(x, 1)$ would define θ in T. (For then if $k \in \theta$, $f(k) = 1$, whence $\vdash_T A(\mathbf{k}, \mathbf{1})$; and if $k \notin \theta$, $f(k) = 0$, whence $\vdash_T \forall y (A(\mathbf{k}, y) \leftrightarrow y = \mathbf{0})$, whence, as $\vdash_Q \mathbf{0} \neq \mathbf{1}$, $\vdash_T -A(\mathbf{k}, \mathbf{1})$.) Thus the characteristic function of θ is not representable in T, and therefore, as T is an extension of Q, not representable in Q either, and hence not recursive. So T is not decidable.

Lemma 4

Q is not decidable.

Proof. Q is a consistent extension of Q.

We can now give another proof of the proposition that first-order logic has no decision procedure, a proof that is rather different from the one given in Chapter 10.

Let L be the theory in L, the language of arithmetic, whose theorems are just the valid sentences in L. All theorems of L are theorems of Q, of course, but as not all of (indeed, none of) the axioms of Q are valid, L is not an extension of Q, and we cannot therefore apply theorem 1. But because Q has only finitely many axioms, we can nonetheless prove that L is not decidable, and hence that there is no effective method for deciding whether or not a first-order sentence is valid.

Theorem 2 (*Church's undecidability theorem*)

L is not decidable.

Proof. Let C be a conjunction of the axioms of Q. Then a sentence A is a theorem of Q iff C implies A, iff $(C \rightarrow A)$ is valid, iff $(C \rightarrow A)$ is a theorem of L. (So, intuitively, a test for validity would yield a test for theoremhood in Q: to decide whether A is a theorem of Q, test $(C \rightarrow A)$ for validity.)

Let q be the gödel number of C. Let f be defined by:

$$f(n) = 1*(q*(39999*(n*2))).$$

f is recursive. If n is the gödel number of A, then $f(n)$ is the gödel number of $(C \rightarrow A)$.

Let λ be the set of gödel numbers of theorems of L. If λ is recursive, then so is $\{n|f(n)\in\lambda\}$. But $\{n|f(n)\in\lambda\}$ is the set of gödel numbers of theorems of Q, which, by lemma 4, is not recursive. Thus λ is not recursive and L is not decidable.

By *arithmetic* we shall understand that theory whose language is L and whose theorems are just the sentences of L that are *true* in the standard interpretation $\mathcal{N}$, in which the domain is the set of all natural numbers, and **o**, $'$, $+$, and $\cdot$ are assigned zero, successor, addition, and multiplication, respectively.

Theorem 3

Arithmetic is not decidable.

Proof. Arithmetic is a consistent extension of Q, and by Theorem 1 no consistent extension of Q is decidable.

Thus, unless Church's thesis is false, there is no effective method for deciding whether an arbitrary sentence in the language of arithmetic is true or false in $\mathcal{N}$. This negative result is in contrast to Presburger's theorem, proved in Chapter 21, that an effective method exists for deciding whether an arbitrary sentence in the language of arithmetic *not containing* '$\cdot$' is true or false (in $\mathcal{N}$).

Theorem 4 (*Tarski's indefinability theorem*)

The set of gödel numbers of sentences true in $\mathcal{N}$ is not definable in arithmetic.

Proof. Since the theorems of arithmetic are just the sentences true in $\mathcal{N}$, Theorem 4 follows from Lemma 3.

As any formula $B(x)$ will be true (in $\mathcal{N}$) of the number k if and only if $B(\mathbf{k})$ is a theorem of arithmetic, another way to put Theorem 4 is to say that there is no formula of the language of arithmetic (with one free variable) which is true of just those natural numbers that are gödel numbers of truths of arithmetic, or, more briefly, 'arithmetical truth is not arithmetically definable'.

Lemma 5

Any recursive set is definable in arithmetic.

Proof. Suppose θ is a recursive set. Then the characteristic function of θ is recursive, and hence representable in Q. As in the proof of Theorem 1, θ is then definable in Q, and hence definable in arithmetic, which is an extension of Q.

Lemma 5 shows that Theorem 4 is at least as strong a result as Theorem 3, as Theorem 3 says that the set of gödel numbers of truths of $\mathcal{N}$ is not recursive. Since the converse of Lemma 5 does not hold (cf. Exercise 3), Theorem 4 is actually stronger than Theorem 3.

A theory T is called *complete* if for every sentence A (in the language of T), either A or $-A$ is a theorem of T. A theory T is consistent and complete, then, iff for any sentence A, exactly one of A and $-A$ is a theorem. Arithmetic is a consistent, complete extension of Q.

A theory T is called *axiomatizable* if there is a decidable subset of T whose consequences (in the language of T) are just the theorems of T. If there is a finite, and hence decidable, subset with this property, the theory is said to be *finitely axiomatizable*. It is clear from the definition of axiomatizability that any decidable theory is axiomatizable; Q is an example of a (finitely) axiomatizable theory that is not decidable.

The version of Gödel's incompleteness theorem that we shall prove is the assertion that there is no complete, consistent, axiomatizable extension of Q. That there is none will follow from Theorem 1 and the proposition (Theorem 5) that any axiomatizable complete theory is decidable.

This last proposition should be confused neither with the statement that every complete decidable theory is axiomatizable, which is trivially true, nor with the statement that every decidable axiomatizable theory is complete, which is false (counterexample: the theory whose non-logical symbols are the sentence letters p and q, and whose theorems are the consequences in this language of p).

Theorem 5

Any axiomatizable complete theory is decidable.

Proof. Let T be any theory whatsoever.. Since the set of symbols that may occur in sentences of the language of T is decidable (as we have assumed earlier), the set of sentences of the language of T is itself decidable, i.e. there exists a Turing machine which, when given a number as input, yields 1 as output iff the number is the gödel number of a sentence of the language of T, and yields 0 otherwise.

Now suppose that T is axiomatizable and complete. If T is inconsistent, then the theorems of T are just the sentences in the language of T, which, we have just noted, form a decidable set. We may therefore suppose that T is consistent also.

Since T is axiomatizable, there is a decidable set S of sentences whose consequences (in the language of T) are just the theorems of T. Let A be a sentence (in the language of T). We shall say that a sentence is *A-interesting* if it is a conditional of which the antecedent is a conjunction of members of S and the consequent is either A or $-A$. Then, A is a theorem of T iff there is a *valid* A-interesting sentence whose consequent is A itself. And, since T is consistent and complete, A is a theorem of T iff $-A$ is not a theorem of T, iff there is *no* valid A-interesting sentence whose consequent is $-A$. Since S is decidable, so is the set of A-interesting sentences (for each sentence A).

We shall show that T is decidable by showing that there exists a Turing machine M which, when given a sentence A of the language of T as input, yields 1 as output iff A is a theorem of T, and yields 0 as output otherwise.

In Chapter 12 we established the existence of a Turing machine M^* which, when given any sentence as input, terminated after a finite number of steps with the production of the words 'yes, valid' iff the given sentence was valid.

Our machine M works by first writing down the number 1 after the input sentence A and then going into a loop consisting of a sequence of *subroutines*. In the nth of these, M writes down those $k(\leqslant n)$ sentences with gödel numbers $\leqslant n$ that are A-interesting and then 'imitates' M^* k times, each time performing n steps in the operation of M^* when given as input (a new) one of the k A-interesting sentences that have been written down. If one or more of these k sentences is shown valid (by the production of the words 'yes, valid') after n such steps, M picks one of them and determines whether its consequent is A or $-A$ (as T is consistent, the case cannot arise in which one sentence has consequent A and another has $-A$), and then yields as output 1 or 0, accordingly. But if not, M erases everything except A and the number after it, to which it adds 1, and then goes into the $n+1$st subroutine.

Since either A or $-A$ is a theorem of T, but not both, there is a valid A-interesting sentence C, with gödel number i, and M^*, when applied to C, will terminate with 'yes, valid' after some finite number of steps, say j. M, therefore, when applied to A, will go into at most $\max(i,j)$ subroutines before yielding 1 or 0 as output, and will yield 1 iff the consequent of C is A. So M yields 1 as output iff A is valid, and yields 0 otherwise. T is therefore decidable.

Theorem 6 (*Gödel's first incompleteness theorem*)

There is no consistent, complete, axiomatizable extension of Q.

Proof. Theorem 6 is an immediate consequence of Theorems 1 and 5.

Corollary

Arithmetic is not axiomatizable.

The import of Gödel's first incompleteness theorem is sometimes expressed in the words 'any sufficiently strong formal theory (or system) of arithmetic is incomplete (if it is consistent)'. A 'formal' theory may be taken to be one whose theorems are deducible via the usual rules of logic from an axiom system. Since an axiom system is here understood to be a set of sentences for which an effective procedure for determining membership exists, and since the usual rules of logic are sound and complete, that is, since all and only the logical consequences of a set of sentences can be deduced from the set by means of the rules, 'formal theory' can be considered synonymous with 'axiomatizable theory'. 'A formal theory of arithmetic' can therefore be taken to be an axiomatizable theory all of whose theorems are truths in some interpretation whose domain is the set of natural numbers and in which those of **0**, $'$, $+$, $\cdot$, $<$, ^{2}, etc. that occur in the theorems have their familiar, standard meanings.

Theorem 6 thus represents a sharpening of the above statement of Gödel's theorem in that it indicates a sufficient condition for 'sufficient strength', *viz.*, *being an extension of Q*. Q, as we have seen, is a rather weak theory (cf. Exercise 14.2), and Theorem 6 is thus a correspondingly strong result. It follows from Theorem 6 that any consistent mathematical theory of which the theorems are just the consequences of some effectively specified set of axioms, and among which are the seven axioms of Q, is incomplete; hence for any interpretation of the language of the theory there will be truths in that interpretation which are not theorems of the theory. And perhaps the most significant consequence of Theorem 6 is what it says about the notions of *truth* (in the standard interpretation for the language of arithmetic) and *theoremhood*, or *provability* (in any particular formal theory): *that they are in no sense the same.*

Exercises

15.1 A formula $B(y)$ is called a truth-predicate for T if for any sentence G of the language of T, $\vdash_T G \leftrightarrow B(\ulcorner G \urcorner)$. Show that if T is a consistent theory in which diag is representable, then there is no truth-predicate for T.

15.2 Show that all functions representable in Q are recursive.

15.3 A set S of natural numbers is called *recursively enumerable* (*r.e.*) if there is a (two-place) recursive relation R such that $S = \{x \mid \exists y\, Rxy\}$. Show that for any set S, S is recursive iff both S and $\bar{S}$ are r.e. (Kleene's theorem). Are all r.e. sets definable in arithmetic? (Yes. Why?) Give some examples of r.e. sets and some examples of non-r.e. sets.

15.4 (*Craig*) Show that a theory T is axiomatizable if T is r.e., i.e. if the set of gödel numbers of members of T is r.e.

15.5 Let $B_1(y)$ and $B_2(y)$ be two formulas of the language of T with y as sole free variable. Show how to construct sentences A_1 and A_2 such that $\vdash_T A_1 \leftrightarrow B_1(\ulcorner A_2 \urcorner)$ and $\vdash_T A_2 \leftrightarrow B_2(\ulcorner A_1 \urcorner)$.

Solution to 15.2 (*Using Church's thesis*)

15.2 Suppose $A(\mathsf{x},y)$ represents f in Q. Since Q is consistent and $\mathbf{m} \neq \mathbf{n}$ is a theorem of Q whenever $m \neq n$, $\vdash_Q \forall y(A(\mathbf{p},y) \leftrightarrow y = \mathbf{m})$ iff $f(\mathsf{p}) = m$. In order to calculate $f(\mathsf{p})$, then, one may use a 'search procedure' similar to the one used in the proof of Theorem 5 to determine for which m the conditional whose antecedent is some fixed conjunction of the axioms of Q and whose consequent is $\forall y(A(\mathbf{p},y) \leftrightarrow y = \mathbf{m})$ is valid. That m – it will be unique – is $f(\mathsf{p})$.

Solution to 15.4 (very tricky)

Suppose that R is a recursive relation and

$$\{x \mid x \text{ is the gödel number of a member of } T\} = \{x \mid \exists y\, Rxy\}.$$

For any sentence A and natural number y, let A^y be the conjunction $(A\&\ldots(A\&A)\ldots)$ of $y+2$ occurrences of A. Thus, e.g.

$$A^2 = (A\&(A\&(A\&A)))$$

and $A^0 = (A\&A)$. Let $U = \{A^y \mid R\,\mathrm{gn}(A)y\}$. If $A \in T$, then for some y, $R\,\mathrm{gn}(A)y$ and $A^y \in U$; and if $A^y \in U$, then $A \in T$. Since A and A^y are equivalent, T and U imply the same sentences, and the set of sentences in the language of T that follow from U is thus T itself. To show that T is axiomatizable, then, we need only show that U is decidable. But U *is* decidable: to decide whether an arbitrary sentence B is in U, we may apply the following effective procedure. Determine whether B is the conjunction $(A\&\ldots(A\&A)\ldots)$ of z occurrences of some sentence A, for some $z \geqslant 2$. If not, $B \notin U$. But if so, find A and z, and let $x = \mathrm{gn}(A)$ and $y = z-2$. Determine whether Rxy. (R is recursive.) If so, $B \in U$; if not, $B \notin U$.

16

Provability predicates and the unprovability of consistency

We learned in the last chapter that no consistent extension of Q was decidable, and that any complete axiomatizable theory was decidable; we concluded that no consistent, axiomatizable extension of Q was complete. In the present chapter we are going to discuss a certain theory, *Elementary Peano Arithmetic*, or Z, as we shall call it (following Hilbert–Bernays). Z's language is L, the language of arithmetic. The *'non-logical' axioms of Z are* $Q1$ through $Q7$, the seven axioms of Q, together with all *induction axioms*: sentences obtained from formulas of L of the form:

$$([A(\mathbf{0}) \,\&\, \forall x(A(x) \rightarrow A(x'))] \rightarrow \forall x A(x))$$

by the prefixing of universal quantifiers. The theorems of Z are the logical consequences in L of the axioms of Q and the induction axioms. ($Q3$ actually follows from an induction axiom. For let

$$A(x) = (x \neq \mathbf{0} \rightarrow \exists y\, x = y').$$

Then $A(\mathbf{0})$ and $\forall x(A(x) \rightarrow A(x'))$ are logical truths; and so

$$\forall x A(x), = Q3, = \forall x(x \neq \mathbf{0} \rightarrow \exists y\, x = y'),$$

is a theorem of Z, and thus its inclusion as a non-logical axiom of Z was not necessary.) Like the axioms of Q, the induction axioms are all true in the standard interpretation $\mathscr{N}$, and hence Z is consistent; Z is even more evidently an axiomatizable extension of Q. Z is thus incomplete.

But though incomplete, because of the presence of the induction axioms, Z is a much more powerful theory than Q: all of the sentences mentioned in Exercise 14.2, for example, are theorems of Z but not of Q. And if Z is supplied with a notion of *proof* (such as the one given below), then the proofs of a great many mathematical theorems, including much (notably) of the theory of numbers, can be 'reproduced' or 'carried out' in Z. Unlike Q, Z is a theory in which large portions of actual mathematics can be adequately formalized. And, as Z is an extension of Q, all recursive functions (sets or relations) are representable (definable) in Z.

In the present chapter we are going to investigate two closely related questions about Z and other theories: whether the consistency of Z is provable in Z, and whether a certain sentence, which may be taken to

assert its own provability (in Z) is in fact true (and thus provable) or false (and thus unprovable). Unlike the question whether a sentence expressing its own unprovability (in Z) is provable or not – any such sentence must be both true and unprovable if everything provable is true – there seems to be no easy way to decide this second question *a priori*.

We shall now introduce, informally, the notion of a *proof of* a sentence in Z. We shall assume, but not prove, this fact: there is an effectively specifiable set of valid sentences, the 'logical' axioms of Z, such that a sentence A is a theorem of Z if and only if there is a finite sequence of sentences (of L) ending with A, each of which is either an axiom of Z (logical or non-) or the 'ponential' of two *earlier* sentences in the sequence. (D is the 'ponential' of C and $(C \rightarrow D)$.) Such a sequence shall be called a *proof in Z of A*.

We shall be a little bit more specific about the 'nature' of finite sequences of sentences: each such sequence is to be identified with the expression consisting of the same sentences in the same order, but separated by commas. As the comma has already been assigned the gödel number 29, every proof in Z acquires its own gödel number in consequence of this identification. The relation Proof, $= \{\langle m, n\rangle \mid m$ is the gödel number of a proof in Z of the sentence with gödel number $n\}$, is thus a recursive relation and is therefore definable in Z.

By 'straightforwardly transcribing' in L the definition of *proof in Z of* just given, making reference to gödel numbers instead of expressions, and utilizing, where necessary, the β-function of Chapter 14, we can construct a formula $\mathrm{Pr}(x, y)$ of L, which not only defines Proof in Z, but possesses certain other important properties as well. In order to describe these, we need a

Definition

$\mathrm{Prov}(y)$ is to be the formula $\exists x\, \mathrm{Pr}(x, y)$.

One important property that $\mathrm{Pr}(x, y)$ and $\mathrm{Prov}(y)$ have is that for any sentences A and C (of L),

(i) if $\vdash_Z A$, then $\vdash_Z \mathrm{Prov}(\ulcorner A \urcorner)$;
(ii) $\vdash_Z \mathrm{Prov}(\ulcorner A \rightarrow C \urcorner) \rightarrow [\mathrm{Prov}(\ulcorner A \urcorner) \rightarrow \mathrm{Prov}(\ulcorner C \urcorner)]$; and
(iii) $\vdash_Z \mathrm{Prov}(\ulcorner A \urcorner) \rightarrow \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner A \urcorner) \urcorner)$.

(Later we shall express this property of $\mathrm{Prov}(y)$ by saying that $\mathrm{Prov}(y)$ is a *provability predicate* for Z. The consequent of the sentence mentioned in (iii) is the result of substituting $\mathbf{k}$ for y in $\mathrm{Prov}(y)$, k being the gödel number of $\text{Prov}(\ulcorner A \urcorner)$, which is itself the result of substituting the numeral for the gödel number of A for y in $\text{Prov}(y)$.) It is also the case that for any sentence A,

(iv) if $\vdash_Z \text{Prov}(\ulcorner A \urcorner)$, then $\vdash_Z A$.

That $\text{Prov}(y)$ satisfies (i) follows directly from the fact that $\text{Pr}(x,y)$ defines Proof in Z. For suppose that $\vdash_Z A$. Then there is a proof of A in Z. Let the gödel number of the proof be m. Then $\vdash_Z \text{Pr}(\mathbf{m}, \ulcorner A \urcorner)$; and so $\vdash_Z \exists x \text{Pr}(x, \ulcorner A \urcorner)$, i.e., $\vdash_Z \text{Prov}(\ulcorner A \urcorner)$.

A proof of C can be obtained from proofs of A and $(A \rightarrow C)$ by writing down the proof of A, then a comma, then the proof of $(A \rightarrow C)$, then another comma, and then C; this argument can be formalized in Z. Its formalization would show that $\text{Prov}(y)$ satisfies (ii).

Showing that $\text{Prov}(y)$ satisfies (iii) is much harder, and we shall only say that it involves showing that the argument that $\text{Prov}(y)$ satisfies (i) (with '$\vdash_Z$' understood as meaning 'provable in Z') can be formalized in Z (for any given A). And showing *that* involves showing, in Z, that for any m, n, if m Proof n, then $\text{Pr}(\mathbf{m}, \mathbf{n})$ is provable in Z.[†]

As for (iv), suppose that $\vdash_Z \text{Prov}(\ulcorner A \urcorner)$, i.e. $\vdash_Z \exists x \text{Pr}(x, \ulcorner A \urcorner)$. Then $\exists x \text{Pr}(x, \ulcorner A \urcorner)$ is true in $\mathcal{N}$, and so for some m, m is the gödel number of a proof in Z of A. So A is provable in Z.

A further important, if non-mathematical, property of $\text{Pr}(x,y)$ is that it can plausibly be regarded as *meaning* or *saying* that the number represented by x is the gödel number of a proof in Z of the sentence whose gödel number is represented by y, or more concisely but somewhat inaccurately, as meaning that x is a proof of y. Not every formula that defines Proof has this property: let $\text{Dr}(x,y)$ be the formula

$$(\text{Pr}(x,y) \,\&\, y \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner).$$

Then, as Z is consistent, $\text{Dr}(x,y)$ also defines Proof, but cannot be said to mean simply that x is a proof of y; it may be regarded as meaning that x is a proof of y, but y is not the sentence $\mathbf{0} = \mathbf{1}$. In brief and roughly, it is because it is the 'straightforward transcription' of the definition of *prcof in Z of* that $\text{Pr}(x,y)$ can be considered to mean what it does.

$\text{Prov}(y)$ can similarly be taken as meaning that (the number represented by) y is (the gödel number of) a provable sentence. In what follows we shall concern ourselves with the questions whether a sentence that asserts its own provability in Z is provable (in Z) or not, and whether the sentence that says that Z is consistent can be proved in Z. We may and shall take these, and other, questions, whose formulations involve notions like *asserting its own provability* or *expressing the non-provability of* $\mathbf{0} = \mathbf{1}$, as questions about various sentences constructed in certain specific ways from $\text{Prov}(y)$, and not from some other formula with some but not all of $\text{Prov}(y)$'s properties.

From now on, T will be an extension of Q, not necessarily consistent, and $B(y)$ will be a formula of the language of T with the sole free variable y. If A is a sentence of the language of T with gödel number n, then $B(\ulcorner A \urcorner)$ is the result of everywhere substituting $\mathbf{n}$ for all free occurrences of y in $B(y)$.

We shall call $B(y)$ a *provability predicate for* T if for all sentences A and C (of the language of T), $B(y)$ meets the following three conditions:

(i) if $\vdash_T A$, then $\vdash_T B(\ulcorner A \urcorner)$;

(ii) $\vdash_T B(\ulcorner A \rightarrow C \urcorner) \rightarrow [B(\ulcorner A \urcorner) \rightarrow B(\ulcorner C \urcorner)]$; and

(iii) $\vdash_T B(\ulcorner A \urcorner) \rightarrow B(\ulcorner B(\ulcorner A \urcorner) \urcorner)$.

$\text{Prov}(y)$ is a provability predicate for Z, but provability predicates need not have very much to do with provability: any formula $S(y)$ that defines $\{n | n$ is the gödel number of a sentence of $L\}$ in Q is a provability predicate for Q. (Note that $\vdash_Q S(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$, but not $\vdash_Q \mathbf{0} = \mathbf{1}$: an analogue of (iv) (above) fails for Q, $S(y)$ and $A = \mathbf{0} = \mathbf{1}$.)

The thought that whatever is provable had better be true might make it surprising that a further condition was not included in the definition of provability predicate, namely,

For every sentence A, $\vdash_T B(\ulcorner A \urcorner) \rightarrow A$.

It will become clear, though, that no provability predicate meets this extra condition unless T is actually inconsistent!

The diagonal lemma of Chapter 15 provided us with a way, when given any formula $B(y)$, of finding a sentence A such that $\vdash_T A \leftrightarrow B(\ulcorner A \urcorner)$. If we take T to be Z and $B(y)$ to be $-\text{Prov}(y)$ or $\text{Prov}(y)$, the diagonal lemma enables us to construct sentences G and H such that

$$\vdash_Z G \leftrightarrow -\text{Prov}(\ulcorner G \urcorner) \quad \text{and} \quad \vdash_Z H \leftrightarrow \text{Prov}(\ulcorner H \urcorner).$$

G and H can be taken as expressing their own unprovability and provability, respectively. In what follows we shall answer the questions whether G and H are true or false, i.e., whether or not $\vdash_Z G$ and $\vdash_Z H$. We shall also answer the question whether or not

$$\vdash_Z -\text{Prov}(\ulcorner \mathbf{0} = \mathbf{1} \urcorner); \qquad -\text{Prov}(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$$

can be held to express the consistency of Z. Let us consider G first.

[†] For a complete account the reader may refer to Hilbert–Bernays, *Grundlagen der Mathematik*, and M. Löb 'Solution of a Problem of Leon Henkin', *Journal of Symbolic Logic* **20** (1955).

If $B(y)$ satisfies condition (i) of the definition of *provability predicate* (whether or not it satisfies (ii) and (iii)), and T is consistent, then if $\vdash_T A \leftrightarrow -B(\ulcorner A \urcorner)$, then not: $\vdash_T A$. For if $\vdash_T A$, then by (i), $\vdash_T B(\ulcorner A \urcorner)$, whence $\vdash_T -A$, and T is inconsistent. Thus since Z is consistent, G is not a theorem of Z (and is thus true). It is of somc interest to observe that if $\mathrm{Pr}^*(x, y)$ defines Proof in Z, $\mathrm{Prov}^*(y) = \exists x\, \mathrm{Pr}^*(x, y)$, and

$$\vdash_Z G^* \leftrightarrow -\mathrm{Prov}^*(\ulcorner G^* \urcorner),$$

then not: $\vdash_Z G^*$; thus that $\mathrm{Prov}(y)$ actually satisfies (ii) and (iii) is irrelevant to whether or not $\vdash_Z G$. It is not irrelevant to whether or not $\vdash_Z H$ and $\vdash_Z -\mathrm{Prov}(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$.

In order to answer these two questions we shall state and prove the principal result of this chapter, Löb's theorem. The proof of Löb's theorem that we shall give resembles a certain 'proof' of the existence of Santa Claus:

Let S be the sentence 'If S is true, Santa exists'.
Assume

(1) S is true.

Then by the logic of identity,

(2) 'If S is true, Santa exists' is true.

From (2) we obtain

(3) If S is true, Santa exists.

And from (1) and (3), we infer

(4) Santa exists.

Having deduced (4) from the assumption (1), we may assert outright

(5) If S is true, Santa exists.

From (5) we obtain

(6) 'If S is true, Santa exists' is true.

And by the logic of identity again we have

(7) S is true.

From (5) and (7) we conclude

(8) Santa exists.

Löb's theorem

If $B(y)$ is a provability predicate for T, then for any sentence A, if $\vdash_T B(\ulcorner A \urcorner) \to A$, then $\vdash_T A$.

Proof. Suppose that

$$\vdash_T B(\ulcorner A \urcorner) \to A. \qquad (1)$$

Let $D(y)$ be the formula $(B(y) \to A)$. The diagonal lemma, applied to $D(y)$, gives us a sentence C such that $\vdash_T C \leftrightarrow D(\ulcorner C \urcorner)$, i.e.,

$$\vdash_T C \leftrightarrow (B(\ulcorner C \urcorner) \to A). \qquad (2)$$

So $$\vdash_T C \to (B(\ulcorner C \urcorner) \to A). \qquad (3)$$

So by virtue of (i) of the definition of provability predicate,

$$\vdash_T B(\ulcorner C \to (B(\ulcorner C \urcorner) \to A)\urcorner). \qquad (4)$$

By virtue of (ii),

$$\vdash_T B(\ulcorner C \to (B(\ulcorner C \urcorner) \to A)\urcorner) \to [B(\ulcorner C \urcorner) \to B(\ulcorner B(\ulcorner C \urcorner) \to A \urcorner)]. \qquad (5)$$

So from (4) and (5) it follows that

$$\vdash_T B(\ulcorner C \urcorner) \to B(\ulcorner B(\ulcorner C \urcorner) \to A \urcorner). \qquad (6)$$

By virtue of (ii) again,

$$\vdash_T B(\ulcorner B(\ulcorner C \urcorner) \to A \urcorner) \to [B(\ulcorner B(\ulcorner C \urcorner)\urcorner) \to B(\ulcorner A \urcorner)]. \qquad (7)$$

So from (6) and (7),

$$\vdash_T B(\ulcorner C \urcorner) \to [B(\ulcorner B(\ulcorner C \urcorner)\urcorner) \to B(\ulcorner A \urcorner)]. \qquad (8)$$

By virtue of (iii),

$$\vdash_T B(\ulcorner C \urcorner) \to B(\ulcorner B(\ulcorner C \urcorner)\urcorner). \qquad (9)$$

So from (8) and (9),

$$\vdash_T B(\ulcorner C \urcorner) \to B(\ulcorner A \urcorner). \qquad (10)$$

From (1) and (10),

$$\vdash_T B(\ulcorner C \urcorner) \to A. \qquad (11)$$

From (2) and (11),

$$\vdash_T C. \qquad (12)$$

By virtue of (i) again,

$$\vdash_T B(\ulcorner C \urcorner). \tag{13}$$

And so finally, from (11) and (13) we have that $\vdash_T A$.

The converse of Löb's theorem is of course trivial. It follows that a necessary and sufficient condition for a sentence A to be a theorem of Z is that $(\text{Prov}(\ulcorner A \urcorner) \rightarrow A)$ be a theorem of Z.

Corollary 1

Suppose that $B(y)$ is a provability predicate for T. Then if $\vdash_T A \leftrightarrow B(\ulcorner A \urcorner)$, then $\vdash_T A$.

Corollary 1 shows us that H, which 'says of itself that it is provable in Z', is provable in Z, and thus true.

Corollary 2 (*Gödel's second incompleteness theorem*)

Suppose that $B(y)$ is a provability predicate for T. Then if T is consistent, not: $\vdash_T -B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$.

Proof. Suppose $\vdash_T -B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$. Then $\vdash_T B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner) \rightarrow \mathbf{0} = \mathbf{1}$. By Löb's theorem, then, $\vdash_T \mathbf{0} = \mathbf{1}$, and as T extends Q, T is inconsistent.

Corollary 2 shows that $-\text{Prov}(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$, which 'expresses the consistency of Z', is *not* provable in Z. ('If Z is consistent', one might be tempted to add. But Z is consistent.)

The connection between Löb's theorem, with

$$T = Z \quad \text{and} \quad B(y) = \text{Prov}(y),$$

and the provability of consistency in Z and its extensions may be stated as follows (this way of viewing the matter was suggested to us by Saul Kripke): Let C be a sentence of L, and let $Z + C$ be the theory whose theorems are the consequences in L of $Z \cup \{C\}$. Then the statement that $\vdash_Z \text{Prov}(\ulcorner A \urcorner) \rightarrow A$ is true if and only if $-\text{Prov}(\ulcorner A \urcorner)$ is a theorem of $Z + -A$; the statement that $\vdash_Z A$ is true if and only if $Z + -A$ is inconsistent. Let us call the sentence $-\text{Prov}(\ulcorner A \urcorner)$ *the consistency of* $Z + -A$. Löb's theorem then amounts to the assertion that if the consistency of $Z + -A$ is a theorem of $Z + -A$, then $Z + -A$ is inconsistent.

A formula $\text{Tr}(x)$ is called a *truth-predicate for* T if for every sentence A of the language of T, $\vdash_T A \leftrightarrow \text{Tr}(\ulcorner A \urcorner)$.

Corollary 3

If T is consistent, then T has no truth-predicate.

Proof. Suppose that $\text{Tr}(x)$ is a truth-predicate for T. Then $\text{Tr}(x)$ is a provability predicate for T. (Why?) Moreover, since $\text{Tr}(x)$ is a truth-predicate, for every sentence A, $\vdash_T \text{Tr}(\ulcorner A \urcorner) \rightarrow A$. It follows from Löb's theorem that every sentence A of the language of T is a theorem of T and thus that T is inconsistent.

Of course, Corollary 3 immediately follows from the diagonal lemma, when applied to the negation of any truth-predicate for T.

We conclude by showing that a formula may satisfy (i) – even with 'if' strengthened to 'iff' – without satisfying (ii) and (iii) and so counting as a provability predicate. For suppose that T is consistent and $B(y)$ is some formula such that for any sentence A, $\vdash_T A$ if and only if $\vdash_T B(\ulcorner A \urcorner)$. Let $D(y)$ be the formula $(B(y) \;\&\; y \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner)$. Then

$$\vdash_T -(B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner) \;\&\; \ulcorner \mathbf{0} = \mathbf{1} \urcorner \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner),$$

i.e., $\vdash_T -D(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$. It follows from Corollary 2 that $D(y)$ is not a provability predicate for T. However, if $\vdash_T A$, then, since T is consistent, A is not the same sentence as $\mathbf{0} = \mathbf{1}$, and so $\ulcorner A \urcorner \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner$, and therefore (as T extends Q), $\vdash_T \ulcorner A \urcorner \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner$; but since $\vdash_T A$, $\vdash_T B(\ulcorner A \urcorner)$, so $\vdash_T (B(\ulcorner A \urcorner) \;\&\; \ulcorner A \urcorner \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner)$, i.e. $\vdash_T D(\ulcorner A \urcorner)$. Conversely, if $\vdash_T D(\ulcorner A \urcorner)$, then $\vdash_T B(\ulcorner A \urcorner)$, whence $\vdash_T A$. So we have shown that although $D(y)$ is not a provability predicate, for any sentence A, $\vdash_T A$ if and only if $\vdash_T D(\ulcorner A \urcorner)$.

Exercises

16.1 Suppose that $B(y)$ is a provability predicate for T and that $D(y)$ is the formula $(B(y) \;\&\; y \neq \ulcorner \mathbf{0} = \mathbf{1} \urcorner)$. Show that $D(y)$ meets condition (iii) but not condition (ii) of the definition of *provability predicate* unless T is inconsistent.

16.2 Suppose that $B(y)$ is a provability predicate for T. Use the existence of a sentence G such that $\vdash_T G \leftrightarrow -B(\ulcorner G \urcorner)$ to construct an 'alternative' proof that if T is consistent, then not: $\vdash_T -B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$. Suggestion: show $\vdash_T B(\ulcorner B(\ulcorner G \urcorner) \urcorner) \rightarrow B(\ulcorner -G \urcorner)$, $\vdash_T B(\ulcorner G \urcorner) \rightarrow B(\ulcorner -G \urcorner)$, and

$$\vdash_T B(\ulcorner G \urcorner) \rightarrow [B(\ulcorner -G \urcorner) \rightarrow B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)].$$

Conclude that if $\vdash_T -B(\ulcorner \mathbf{0} = \mathbf{1} \urcorner)$, then $\vdash_T -B(\ulcorner G \urcorner)$, $\vdash_T G$, $\vdash_T B(\ulcorner G \urcorner)$.

16.3 Suppose that T is consistent and that for every sentence A of the language of T, $\vdash_T A$ if and only if $\vdash_T B(\ulcorner A \urcorner)$. Let $D(y)$ be the formula $(B(y) \& y \neq y)$. Then the diagonal lemma supplies a C such that

$$\vdash_T C \leftrightarrow D(\ulcorner C \urcorner).$$

Let $E(y)$ be the formula $(B(y) \& y \neq \ulcorner C \urcorner)$. Show that for every sentence A, $\vdash_T A$ if and only if $\vdash_T E(\ulcorner A \urcorner)$, $\vdash_T C \leftrightarrow E(\ulcorner C \urcorner)$, but not: $\vdash_T C$. Explain why there is no conflict with Corollary 1. What happens if $D^*(y)$ is the formula $(B(y) \mathrm{v}\, y = y)$, C^* is such that $\vdash_T C^* \leftrightarrow D^*(\ulcorner C^* \urcorner)$, and $E^*(y)$ is the formula $(B(y) \mathrm{v}\, y = \ulcorner C^* \urcorner)$?

16.4 Explain why any truth predicate for T is a provability predicate for T. Explain why T is inconsistent if some provability predicate $B(y)$ is such that $\vdash_T B(\ulcorner A \urcorner) \rightarrow A$, for every sentence A.

16.5 Suppose that $B(y)$ is a provability predicate for T and that $\vdash_T B(\ulcorner A \urcorner) \rightarrow C$ and $\vdash_T B(\ulcorner C \urcorner) \rightarrow A$. Show that $\vdash_T A$ and $\vdash_T C$.

16.6 Show that if $B(y)$ is a provability predicate for T, then

$$\vdash_T B(\ulcorner (B(\ulcorner A \urcorner) \rightarrow A) \urcorner) \rightarrow B(\ulcorner A \urcorner).$$

(*Hint*: show $\vdash_T B(\ulcorner L \urcorner) \rightarrow L$, where $L = B(\ulcorner (B(\ulcorner A \urcorner) \rightarrow A) \urcorner) \rightarrow B(\ulcorner A \urcorner)$.)

17
Non-standard models of arithmetic

We are now going to take up the topic of models of arithmetic. We know that there is at least one model of arithmetic, $\mathscr{N}$, the standard interpretation for the language L of arithmetic. Of course, that $\mathscr{N}$ is a model of arithmetic is true by definition: arithmetic is just the set of sentences of L that are true in $\mathscr{N}$. We shall enquire whether there is any sense at all in which $\mathscr{N}$ is the *unique* model of arithmetic.

Of course, no satisfiable set of sentences has exactly one model, literally speaking: given any model, one can construct another that is *isomorphic* but not identical, to it by 'replacing' some element of the domain by another object that is nowhere in the domain.

The definition of isomorphism of interpretations is this: An interpretation $\mathscr{I}$ *is isomorphic to* an interpretation $\mathscr{J}$ if

(1) $\mathscr{I}$ and $\mathscr{J}$ are interpretations of the same languages;

(2) $\mathscr{I}$ and $\mathscr{J}$ assign the same sentence letters the same truth-values; and

(3) there is a one–one function $\mathfrak{h}$ with domain the domain of $\mathscr{I}$ and range the domain of $\mathscr{J}$ such that
(*a*) if $\mathscr{I}$ assigns a name the designation d, then $\mathscr{J}$ assigns it $\mathfrak{h}(d)$;
(*b*) if $\mathscr{I}$ assigns a function symbol the n-place function $\mathfrak{f}$, then $\mathscr{J}$ assigns it the function $\mathfrak{g}$ such that for any $d_1, \ldots, d_n, d$ in the domain of $\mathscr{I}$, $\mathfrak{f}(d_1, \ldots, d_n) = d$ iff $\mathfrak{g}\,(\mathfrak{h}(d_1), \ldots, \mathfrak{h}(d_n)) = \mathfrak{h}(d)$;
(*c*) if $\mathscr{I}$ assigns a predicate letter the n-place characteristic function ϕ, then $\mathscr{J}$ assigns it the characteristic function ψ such that for any $d_1, \ldots, d_n$ in the domain of $\mathscr{I}$, $\phi(d_1, \ldots, d_n) = \psi(\mathfrak{h}(d_1), \ldots, \mathfrak{h}(d_n))$.

We leave it to the reader to verify that the relation *is isomorphic to* is an equivalence relation, and that the same sentences are true in isomorphic interpretations.

If T is a consistent theory of which any two models are *isomorphic*, then T comes as close to the ideal of having exactly one model as any theory could reasonably be expected to. Such theories are said to 'characterize' their models 'up to isomorphism' and to have 'essentially' one model. A definition, then: T is a *categorical* theory if any two models of T (that are interpretations of T's language) are isomorphic. So we might wonder whether arithmetic is categorical.