27

Modal Logic and Provability

Modal logic extends 'classical' logic by adding new logical operators $\Box$ and $\Diamond$ for 'necessity' and 'possibility'. Section 27.1 is an exposition of the rudiments of (sentential) modal logic. Section 27.2 indicates how a particular system of modal logic **GL** *is related to the kinds of questions about provability in* **P** *we considered in Chapters 17 and 18. This connection motivates the closer examination of* **GL** *then undertaken in section 27.3.*

27.1 Modal Logic

Introductory textbooks in logic devote considerable attention to a part of logic we have not given separate consideration: *sentential logic*. In this part of logic, the *only* nonlogical symbols are an enumerable infinity of *sentence letters*, and the only logical operators are negation, conjunction, and disjunction: $\sim$, &, $\vee$. Alternatively, the operators may be taken to be the constant false ($\perp$) and the conditional ($\rightarrow$). The syntax of sentential logic is very simple: sentence letters are sentences, the constant $\perp$ is a sentence, and if A and B are sentences, so is $(A \rightarrow B)$.

The semantics is also simple: an interpretation is simply an assignment ω of truth values, true (represented by 1) or false (represented by 0), to the sentence letters. The valuation is extended to formulas by letting $\omega(\perp) = 0$, and letting $\omega(A \rightarrow B) = 1$ if and only if, if $\omega(A) = 1$, then $\omega(B) = 1$. In other words, $\omega(A \rightarrow B) = 1$ if $\omega(A) = 0$ or $\omega(B) = 1$ or both, and $\omega(A \rightarrow B) = 0$ if $\omega(A) = 1$ and $\omega(B) = 0$. $\sim A$ may be considered an abbreviation for $(A \rightarrow \perp)$, which works out to be true if and only if A is false. $(A \,\&\, B)$ may similarly be taken to be an abbreviation for $\sim(A \rightarrow \sim B)$, which works out to be true if and only if A and B are both true, and $(A \vee B)$ may be taken to be an abbreviation for $(\sim A \rightarrow B)$.

Validity and implication are defined in terms of interpretations: a sentence D is implied by a set of sentences Γ if it is true in every interpretation in which all sentences in Γ are true, and D is valid if it is true in all interpretations. It is decidable whether a given sentence D is valid, since whether D comes out true on an interpretation ω depends only on the values ω assigns to the finitely many sentence letters that occur in D. If there are only k of these, this means that only a finite number of interpretations, namely 2^k of them, need to be checked to see if they make D true. Similar remarks apply to implication.

What is done in introductory textbooks that we have not done here is to work out many particular examples of valid and invalid sentences, and implications and nonimplications among sentences. We are simply going to presume a certain facility with recognizing sentential validity and implication.

Modal sentential logic adds to the apparatus of ordinary or 'classical' sentential logic one more logical operator, the box $\Box$, read 'necessarily' or 'it must be the case that'. One more clause is added to the definition of sentence: if A is a sentence, so is $\Box A$. The diamond $\Diamond$, read 'possibly' or 'it may be the case that', is treated as an abbreviation: $\Diamond A$ abbreviates $\sim\Box\sim A$.

A modal sentence is said to be a *tautology* if it can be obtained from a valid sentence of nonmodal sentential logic by substituting modal sentences for sentence letters. Thus, since $p \vee \sim p$ is valid for any sentence letter p, $A \vee \sim A$ is a tautology for any modal sentence A. Analogously, *tautological consequence* for modal logic is definable in terms of implication for nonmodal sentential logic. Thus since q is implied by p and $p \rightarrow q$ for any sentence letters p and q, B is a tautologous consequence of A and $A \rightarrow B$ for any modal sentences A and B. The inference from A and $A \rightarrow B$ to B is traditionally called *modus ponens*.

There is no single accepted view as to what modal sentences are to be considered modally valid, beyond tautologies. Rather, there are a variety of systems of modal logic, each with its own notion of a sentence being demonstrable.

The *minimal* system of modal sentential logic, **K**, may be described as follows. The *axioms* of **K** include all tautologies, and all sentences of the form

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

The *rules* of **K** allow one to pass from earlier sentences to any sentence that is a tautologous consequence of them, and to pass

$$\text{from } A \text{ to } \Box A.$$

The latter rule is called the rule of *necessitation*. A *demonstration* in **K** is a sequence of sentences, each of which either is an axiom or follows from earlier ones by a rule. A sentence is then *demonstrable* in **K**, or a *theorem* of **K**, if it is the last sentence of some demonstration. Given a finite set $\Gamma = \{C_1, \ldots, C_n\}$, we write $\wedge C$ for the conjunction of all its members, and say Γ is *inconsistent* if $\sim\wedge C$ is a theorem. We say a sentence D is *deducible* from Γ if $\wedge C \rightarrow D$ is a theorem. The usual relationships hold.

Stronger systems can be obtained by adding additional classes of sentences as axioms, resulting in a larger class of theorems. The following are among the candidates:

(A1) $\Box A \rightarrow A$

(A2) $A \rightarrow \Box\Diamond A$

(A3) $\Box A \rightarrow \Box\Box A$

(A4) $\Box(\Box A \rightarrow A) \rightarrow \Box A$.

For any system **S** we write $\vdash_{\mathbf{S}} A$ to mean that A is a theorem of **S**.

There is a notion of *interpretation* or *model* for **K**. We are going to be interested only in *finite* models, so we build finiteness into the definition. A *model* for **K** will be a triple $\mathcal{W} = (W, >, \omega)$, where W is a nonempty finite set, $>$ a two-place relation on it, and ω a valuation or assignment of truth values true or false (represented by 1 or 0) not to sentence letters but to *pairs* (w, p) consisting of an element w of W and a sentence letter p. The notion $\mathcal{W}, w \models A$ of a sentence A being *true* in a model $\mathcal{W}$ and an element w is defined by induction on complexity. The clauses are as follows:

$\mathcal{W}, w \models p$ for p a sentence letter	iff	$\omega(w, p) = 1$
not $\mathcal{W}, w \models \bot$		
$\mathcal{W}, w \models (A \to B)$	iff	not $\mathcal{W}, w \models A$ *or* $\mathcal{W}, w \models B$
$\mathcal{W}, w \models \Box A$	iff	$\mathcal{W}, v \models A$ for all $v < w$.

(We have written $v < w$ for $w > v$.) Note that the clauses for $\bot$ and $\to$ are just like those for nonmodal sentential logic. We say a sentence A is *valid* in the model $\mathcal{W}$ if $\mathcal{W}, w \models A$ for all w in W.

Stronger notions of model of can be obtained by imposing conditions that the relation $>$ must fulfill, resulting in a smaller class of models. The following are among the candidates.

$(W1)$	*Reflexivity*:	for all w,	$w > w$
$(W2)$	*Symmetry*:	for all w and v,	if $w > v$, then $v > w$
$(W3)$	*Transitivity*:	for all w, v, and u,	if $w > v > u$, then $w > u$
$(W4)$	*Irreflexivity*:	for all w,	not $w > w$.

(We have written $w > v > u$ for $w > v$ and $v > u$.) For any class Σ of models, we say A is *valid* in Σ, and write $\models_\Sigma A$, if A is valid in all $\mathcal{W}$ in Σ.

Let **S** be a system obtained by adding axioms and Σ a class obtained by imposing conditions on $>$. If whenever $\vdash_{\mathbf{S}} A$ we have $\models_\Sigma A$, we say **S** is *sound* for Σ. If whenever $\models_\Sigma A$ we have $\vdash_{\mathbf{S}} A$, we say **S** is *complete* for Σ. A soundness and completeness theorem relating the system **S** to a class of models Σ generally tells us that the (set of theorems of) the system **S** is decidable: given a sentence A, to determine whether or not A is a theorem, one can simultaneously run through all demonstrations and through all finite models, until one finds either a demonstration of A or a model of $\sim A$. A large class of such soundness and completeness theorems are known, of which we state the most basic as our first theorem.

27.1 Theorem (Kripke soundness and completeness theorems). Let **S** be obtained by adding to **K** a subset of {(A1), (A2), (A3)}. Let Σ be obtained by imposing on $<_W$ the corresponding subset of {(W1), (W2), (W3)}. Then **S** is sound and complete for Σ.

Since there are eight possible subsets, we have eight theorems here. We are going to leave most of them to the reader, and give proofs for just two: the case of the empty set, and the case of the set {(A3)} corresponding to {(W3)}: **K** is sound and complete for the class of all models, and $\boldsymbol{K} + (A3)$ is sound and complete for the class of transitive models. Before launching into the proofs we need a couple of simple facts.

27.2 Lemma. For any extension **S** of **K**, if $\vdash_{\mathbf{S}} A \to B$, then $\vdash_{\mathbf{S}} \Box A \to \Box B$.

Proof: Suppose we have a proof of $A \to B$. Then we can then extend it as follows:

(1)	$A \to B$	G
(2)	$\Box(A \to B)$	N(1)
(3)	$\Box(A \to B) \to (\Box A \to \Box B)$	A
(4)	$\Box A \to \Box B$	T(2), (3)

The annotations mean: G[iven], [by] N[ecessitation from step] (1), A[xiom], and T[autological consequence of steps] (2), (3).

27.3 Lemma. $\vdash_{\mathbf{K}} (\Box A \,\&\, \Box B) \leftrightarrow \Box(A \,\&\, B)$, and similarly for more conjuncts.

Proof:

(1)	$(A \,\&\, B) \to A$	T
(2)	$\Box(A \,\&\, B) \to \Box A$	25.2(1)
(3)	$\Box(A \,\&\, B) \to \Box B$	S(2)
(4)	$A \to (B \to (A \,\&\, B))$	T
(5)	$\Box A \to \Box(B \to (A \,\&\, B))$	25.2(4)
(6)	$\Box(B \to (A \,\&\, B)) \to (\Box B \to \Box(A \,\&\, B))$	A
(7)	$(\Box A \,\&\, \Box B) \leftrightarrow \Box(A \,\&\, B)$	T(2), (3), (5), (6)

The first three annotations mean: T[autology], [by Lemma] 25.2 [from] (1), and S[imilar to] (2).

Proof of Theorem 27.1: There are four assertions to be proved.

K *is sound for the class of all models*. Let $\mathcal{W}$ be any model, and write $w \models A$ for $\mathcal{W}$, $w \models A$. It will be enough to show that if A is an axiom, then for all w we have $w \models A$, and that if A follows by a rule from $B_1, \ldots, B_n$, and for all w we have $w \models B_i$ for each i, then for all w we have $w \models A$.

Axioms. If A is tautologous, the clauses of the definition of $\models$ for $\bot$ and $\to$ guarantee that $w \models A$. As for axioms of the other kind, if $w \models \Box(A \to B)$ and $w \models \Box A$, then for any $v < w$, $v \models A \to B$ and $v \models A$. Hence $v \models B$ for any $v < w$, and $w \models \Box B$. So $w \models \Box(A \to B) \to (\Box A \to \Box B)$.

Rules. If A is a tautologous consequence of the B_i and $w \models B_i$ for each i, then again the clauses of the definition of $\models$ for $\bot$ and $\to$ guarantee that $w \models A$. For the other rule, if $w \models A$ for all w, then *a fortiori* for any w and any $v < w$, we have $v \models A$. So $w \models \Box A$.

K *is complete for the class of all models*. Suppose A is not a theorem. We construct a model in which A is not valid. We call a sentence a *formula* if it is either a subsentence of A or the negation of one. We call a consistent set of formulas *maximal* if for every formula B it contains one of every pair of formulas B, $\sim B$. First note that $\{\sim A\}$ is consistent: otherwise $\sim\sim A$ is a theorem, and hence A is, as a tautologous consequence. Further, note that every consistent set Γ is a subset of some maximal set: $\wedge\Gamma$ is equivalent to some nonempty disjunction each of whose conjuncts is a conjunction of formulas that contains the members of Γ and contains every formula exactly once, plain or negated. Further, note that a maximal set contains any formula

deducible from it: otherwise it would contain the *negation* of that formula; but a set that contains the negation of a formula deducible from it is inconsistent.

Let W be the set of all maximal sets. W is not empty, since $\{\sim A\}$ is consistent and therefore a subset of some maximal set. W is finite: if there are only k subsentences of A, there are at most 2^k maximal sets. Define a relation $>$ on W by letting $w > v$ if and only if whenever a formula $\Box A$ is in w, the formula A is in v. Finally, for w in W and sentence letter p, let $\omega(w, p) = 1$ if p is in w, and $\omega(w, p) = 0$ if not. Let $\mathcal{W} = (W, >, \omega)$. We are going to show by induction on complexity that for any w in W and any formula B we have $\mathcal{W}, w \models B$ if and only if B is in w. Since there is a w containing $\sim A$ rather than A, it follows that A is not valid in $\mathcal{W}$.

For the base step, if B is a sentence letter p, then p is in w iff $\omega(w, p) = 1$ iff $w \models p$. If B is $\bot$, then $\bot$ is not in w, since w is consistent, and also it is not the case that $w \models \bot$. For the induction step, if B is $C \rightarrow D$, then C and D are subsentences of A, and $\sim B \leftrightarrow (C \,\&\, \sim D)$ is a theorem, being tautologous. Thus B is not in w iff (by maximality) $\sim B$ is in w, iff C and $\sim D$ are in w, iff (by the induction hypothesis) $w \models C$ and not $w \models D$, iff not $w \models C \rightarrow D$. If B is $\Box C$, the induction hypothesis is that for any v, $v \models C$ iff C is in v. We want to show that $w \models \Box C$ iff $\Box C$ is in w. For the 'if' direction, suppose $\Box C$ is in w. Then for any $v < w$, C is in v and so $v \models C$. It follows that $w \models \Box C$.

For the 'only if' direction, suppose $w \models \Box C$. Let

$$V = \{D_1, \ldots, D_m, \sim C\}$$

where the $\Box D_i$ for $1 \leq i \leq m$ are all the formulas in w that begin with $\Box$. Is V consistent? If it is, then it is contained in some maximal v. Since all D_i are in v, we have $v < w$. Since $\sim C$ is in v, not $v \models C$, which is impossible, since $w \models \Box C$. So V is inconsistent, and it follows that

$$(D_1 \,\&\, \cdots \,\&\, D_m) \rightarrow C$$

is a theorem. By Lemma 27.2,

$$\Box(D_1 \,\&\, \cdots \,\&\, D_m) \rightarrow \Box C$$

is a theorem, and so by Lemma 27.3,

$$(\Box D_1 \,\&\, \cdots \,\&\, \Box D_m) \rightarrow \Box C$$

is a theorem. Hence, since each $\Box D_i$ is in w, $\Box C$ is in w.

K + (A3) *is sound for transitive models.* If $w \models \Box A$, then for any $v < w$ it is the case that for any $u < v$ we have by transitivity $u < w$, and so $u \models A$. Thus $v \models \Box A$ for any $v < w$, and $w \models \Box\Box A$. Thus $w \models \Box A \rightarrow \Box\Box A$.

K + (A3) *is complete for transitive models.* The construction used to prove **K** complete for the class of all models needs to be modified. Define $w > v$ if and only if whenever a formula $\Box B$ is in w, the formulas $\Box B$ are both B in v. Then $>$ will be transitive. For if $w > v > u$, then whenever $\Box A$ is in w, $\Box A$ and A will be in v, and since the former is in v, both will also by in u, so $w > u$.

The only other part of the proof that needs modification is the proof that if $w \models \Box C$, then $\Box C$ is in w. So suppose $w \models \Box C$, and let

$$V = \{\Box D_1, D_1, \ldots, \Box D_m, D_m, \sim C\}$$

where the $\Box D_i$ are all the formulas in w that begin with $\Box$. If V is consistent and v is a maximal set containing it, then $w > v$ and $v \models \sim C$, which is impossible. It follows that

$$\Box D_1 \,\&\, D_1 \,\&\, \cdots \,\&\, \Box D_m \,\&\, D_m \rightarrow C$$
$$\Box(\Box D_1 \,\&\, D_1 \,\&\, \cdots \,\&\, \Box D_m \,\&\, D_m) \rightarrow \Box C$$
$$(\Box\Box D_1 \,\&\, \Box D_1 \,\&\, \cdots \,\&\, \Box\Box D_m \,\&\, \Box D_m) \rightarrow \Box C$$

are theorems, and hence any tautologous consequence of the last of these and the axioms $\Box D_i \rightarrow \Box\Box D_i$ is a theorem, and this includes

$$(\Box D_1 \,\&\, \cdots \,\&\, \Box D_m) \rightarrow \Box C$$

from which it follows that $w \models \Box C$.

Besides its use in proving decidability, the preceding theorem makes it possible to prove syntactic results by semantic arguments. Let us give three illustrations. In both the first and the second, A and B are arbitrary sentences, q a sentence letter not contained in either, $F(q)$ any sentence, and $F(A)$ and $F(B)$ the results of substituting A and B respectively for any and all occurrences of q in F. In the second and third, $\blacksquare A$ abbreviates $\Box A \,\&\, A$. In the third, $\bullet A$ is the result of replacing $\Box$ by $\blacksquare$ throughout A.

27.4 Proposition. If $\vdash_{\mathbf{K}} A \leftrightarrow B$, then $\vdash_{\mathbf{K}} F(A) \leftrightarrow F(B)$.

27.5 Proposition. $\vdash_{\mathbf{K}+(\mathrm{A3})} \blacksquare(A \leftrightarrow B) \rightarrow \blacksquare(F(A) \leftrightarrow F(B))$.

27.6 Proposition. If $\vdash_{\mathbf{K}+(\mathrm{A1})+(\mathrm{A3})} A$, then $\vdash_{\mathbf{K}+(\mathrm{A3})} \bullet A$.

Proof: For Proposition 27.4, it is easily seen (by induction on complexity of F) that if $\mathcal{W} = (W, >, \omega)$ and we let $\mathcal{W}' = (W, >, \omega')$, where ω' is like ω except that for all w

$$\omega'(w, q) = 1 \quad \text{if and only} \quad \text{if } \mathcal{W}, w \models A$$

then for all w, we have

$$\mathcal{W}, w \models F(A) \quad \text{if and only} \quad \text{if } \mathcal{W}', w \models F(q).$$

But if $\vdash_{\mathbf{K}} A \leftrightarrow B$, then by soundness for all w we have

$$\mathcal{W}, w \models A \quad \text{if and only} \quad \text{if } \mathcal{W}, w \models B$$

and hence

$$\mathcal{W}, w \models F(B) \quad \text{if and only} \quad \text{if } \mathcal{W}', w \models F(q)$$
$$\mathcal{W}, w \models F(A) \quad \text{if and only} \quad \text{if } \mathcal{W}, w \models F(B).$$

So by completeness we have $\vdash_{\mathbf{K}} F(A) \leftrightarrow F(B)$.

For Proposition 27.5, it is easily seen (by induction on complexity of A) that since each clause in the definition of truth at w mentions only w and those v with $w > v$, for any $\mathcal{W} = (W, >, \omega)$ and any w in W, whether $\mathcal{W}, w \models A$ depends only on the values of $\omega(v, p)$ for those v such that there is a sequence

$$w = w_0 > w_1 > \cdots > w_n = v.$$

If $>$ is transitive, these are simply those v with $w \geq v$ (that is, $w = v$ or $w > v$). Thus for any transitive model $(W, >, \omega)$ and any w, letting $W_w = \{v\colon w \geq v\}$ and $\mathcal{W}_w = (W_w, >, \omega)$, we have

$$\mathcal{W}, w \models A \quad \text{if and only if} \quad \mathcal{W}_w, w \models A.$$

Now

$$\mathcal{W}, w \models \boxdot C \quad \text{if and only if} \quad \text{for all } v \leq w \quad \text{we have} \quad \mathcal{W}, v \models C.$$

Thus if $\mathcal{W}, w \models \boxdot(A \leftrightarrow B)$, then $\mathcal{W}_w, v \models A \leftrightarrow B$ for all v in W_w. Then, arguing as in the proof of Proposition 27.4, we have $\mathcal{W}_w, v \models F(A) \leftrightarrow F(B)$ for all such v, and so $\mathcal{W}, w \models \boxdot(F(A) \leftrightarrow F(B))$. This shows

$$\mathcal{W}, w \models \boxdot(A \leftrightarrow B) \rightarrow \boxdot(F(A) \leftrightarrow F(B))$$

for all transitive $\mathcal{W}$ and all w, from which the conclusion of the proposition follows by soundness and completeness.

For Proposition 27.6, for any model $\mathcal{W} = (W, >, \omega)$, let $\bullet\mathcal{W} = (W, \geq, \omega)$. It is easily seen (by induction on complexity) that for any A and any w in W

$$\mathcal{W}, w \models A \quad \text{if and only if} \quad \bullet\mathcal{W}, w \models \bullet A.$$

$\bullet\mathcal{W}$ is always reflexive, is the same as $\mathcal{W}$ if $\mathcal{W}$ was already reflexive, and is transitive if and only if $\mathcal{W}$ was transitive. It follows that A is valid in all transitive models if and only if $\bullet A$ is valid in all reflexive transitive models. The conclusion of the proposition follows by soundness and completeness.

The conclusion of Proposition 27.4 actually applies to *any system containing* **K** in place of **K**, and the conclusions of Propositions 27.5 and 27.6 to *any system containing* **K** + (A3) in place of **K** + (A3). We are going to be especially interested in the system **GL** = **K** + (A3) + (A4). The soundness and completeness theorems for **GL** are a little tricky to prove, and require one more preliminary lemma.

27.7 Lemma. If $\vdash_{\mathbf{GL}} (\Box A \,\&\, A \,\&\, \Box B \,\&\, B \,\&\, \Box C) \rightarrow C$, then $\vdash_{\mathbf{GL}} (\Box A \,\&\, \Box B) \rightarrow \Box C$, and similarly for any number of conjuncts.

Proof: The hypothesis of the lemma yields

$$\vdash_{\mathbf{GL}} (\Box A \,\&\, A \,\&\, \Box B \,\&\, B) \rightarrow (\Box C \rightarrow C).$$

Then, as in the proof of the completeness of **K** + (A3) for transitive models, we get

$$\vdash_{\mathbf{GL}} (\Box A \,\&\, \Box B) \rightarrow \Box(\Box C \rightarrow C).$$

From this and the axiom $\Box(\Box C \to C) \to \Box C$ we get as a tautologous consequence the conclusion of the lemma.

27.8 Theorem (Segerberg soundness and completeness theorems). **GL** is sound and complete for transitive, irreflexive models.

Proof: *Soundness.* We need only show, in addition to what has been shown in the proof of the soundness of $\mathbf{K} +$ (A3) for transitive models, that if a model is also irreflexive, then $w \models \Box(\Box B \to B) \to \Box B$ for any w. To show this we need a notion of *rank*.

First note that if $>$ is a transitive, irreflexive relation on a nonempty set W, then whenever $w_0 > w_1 > \cdots > w_m$, by transitivity we have $w_i > w_j$ whenever $i < j$, and hence by irreflexivity $w_i \neq w_j$ whenever $i \neq j$. Thus if W has only m elements, we can never have $w_0 > w_1 > \cdots > w_m$. Thus in any transitive, irreflexive model, there is for any w a greatest natural number k for which there exists elements $w = w_0 > \cdots > w_k$. We call this k the *rank* $\mathrm{rk}(w)$ of w. If there is no $v < w$, then $\mathrm{rk}(w) = 0$. If $v < w$, then $\mathrm{rk}(v) < \mathrm{rk}(w)$. And if $j < \mathrm{rk}(w)$, then there is an element $v < w$ with $\mathrm{rk}(v) = j$. (If $w = w_0 > \cdots > w_{\mathrm{rk}(w)}$, then $w_{\mathrm{rk}(w)-j}$ is such a v.)

Now suppose $w \models \Box(\Box B \to B)$ but not $w \models \Box B$. Then there is some $v < w$ such that not $v \models B$. Take such a v *of lowest possible rank*. Then for all $u < v$, by transitivity $u < w$, and since $\mathrm{rk}(u) < \mathrm{rk}(v)$, $u \models B$. This shows $v \models \Box B$, and since not $v \models B$, not $v \models \Box B \to B$. But that is impossible, since $v < w$ and $w \models \Box(\Box B \to B)$. Thus if $w \models \Box(\Box B \to B)$ then $w \models \Box B$, so for all w, $w \models \Box(\Box B \to B) \to \Box B$.

Completeness. We modify the proof of the completeness of $\mathbf{K} +$ (A3) by letting W be not the set of all maximal w, but only of those for which not $w > w$. This makes the model irreflexive.

The only other part of the proof that needs modification is the proof that if $w \models \Box C$, then $\Box C$ is in w. So suppose $w \models \Box C$, and let

$$V = \{\Box D_1, D_1, \ldots, \Box D_m, D_m, \Box C, {\sim}C\}$$

where the $\Box D_i$ are all the formulas in w that begin with $\Box$. If V is consistent and v is a maximal set containing it, then since $\Box C$ is in v but C cannot be in v, we have not $v > v$, and v is in W. Also $w > v$ and $v \models {\sim}C$, which is impossible. It follows that

$$\Box D_1 \,\&\, D_1 \,\&\, \cdots \,\&\, \Box D_m \,\&\, D_m \,\&\, \Box C \to C$$

is a theorem, and hence by the preceding lemma so is

$$(\Box D_1 \,\&\, \cdots \,\&\, \Box D_m) \to \Box C$$

from which it follows that $w \models \Box C$.

27.2 The Logic of Provability

Let us begin by explaining why the system **GL** is of special interest in connection with the matters with which we have been concerned through most of this book. Let L

be the language of arithmetic, and ϕ a function assigning to sentence letters sentences of L. We associate to any modal sentence A a sentence A^ϕ of L as follows:

$$\begin{aligned} p^\phi &= \phi(p) \qquad \text{for } p \text{ a sentence letter} \\ \bot^\phi &= \mathbf{0} = \mathbf{1} \\ (B \to C)^\phi &= B^\phi \to C^\phi \\ (\Box B)^\phi &= \mathrm{Prv}(\ulcorner B^\phi \urcorner) \end{aligned}$$

where Prv is a provability predicate for **P**, in the sense of chapter 18. Then we have the following relationship between **GL** and **P**:

27.9 Theorem (Arithmetical soundness theorem). If $\vdash_{\mathbf{GL}} A$, then for all ϕ, $\vdash_{\mathbf{P}} A^\phi$.

Proof: Fix any ϕ. It is sufficient to show that $\vdash_{\mathbf{P}} A^\phi$ for each axiom of **GL**, and that if B follows by rules of **GL** from $A_1, \ldots, A_m$ and $\vdash_{\mathbf{P}} A_i^\phi$ for $1 \le i \le m$, then $\vdash_{\mathbf{P}} B^\phi$. This is immediate for a tautologous axioms, and for the rule permitting passage to tautologous consequences, so we need only consider the three kinds of modal axioms, and the one modal rule, necessitation. For necessitation, what we want to show is that if $\vdash_{\mathbf{P}} B^\phi$, then $\vdash_{\mathbf{P}} (\Box B)^\phi$, which is to say $\vdash_{\mathbf{P}} \mathrm{Prv}(\ulcorner B^\phi \urcorner)$. But this is precisely property (P1) in the definition of a provability predicate in Chapter 18 (Lemma 18.2). The axioms $\Box(B \to C) \to (\Box B \to \Box C)$ and $\Box B \to \Box\Box B$ correspond in the same way to the remaining properties (P2) and (P3) in that definition.

It remains to show that $\vdash_{\mathbf{P}} A^\phi$ where A is an axiom of the form

$$\Box(\Box B \to B) \to \Box B.$$

By Löb's theorem it suffices to show $\vdash_{\mathbf{P}} \mathrm{Prv}(\ulcorner A^\phi \urcorner) \to A^\phi$. To this end, write S for B^ϕ, so that A^ϕ is

$$\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \to \mathrm{Prv}(\ulcorner S \urcorner).$$

By (P2)

$$\mathrm{Prv}(\ulcorner A^\phi \urcorner) \to [\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \urcorner) \to \mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \urcorner)]$$

$$\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \to [\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \urcorner) \to \mathrm{Prv}(\ulcorner S \urcorner)]$$

are theorems of **P**, and by (P3)

$$\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \to \mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \urcorner)$$

is also a theorem of **P**. And therefore

$$\mathrm{Prv}(\ulcorner A^\phi \urcorner) \to [\mathrm{Prv}(\ulcorner \mathrm{Prv}(\ulcorner S \urcorner) \to S \urcorner) \to \mathrm{Prv}(\ulcorner S \urcorner)]$$

which is to say $\mathrm{Prv}(\ulcorner A^\phi \urcorner) \to A^\phi$, being a tautological consequences of these three sentences, is a theorem of **P** as required.

The converse of Theorem 27.9 is the *Solovay completeness theorem*: if for all ϕ, $\vdash_{\mathbf{P}} A^\phi$, then $\vdash_{\mathbf{GL}} A$. The proof of this result, which will not be needed in what follows, is beyond the scope of a book such as this.

Theorem 27.9 enables us to establish results about provability in **P** by establishing results about **GL**. The remainder of this section will be devoted to the statement of two results about **GL**, the *De Iongh–Sambin fixed point theorem* and a *normal form theorem for letterless sentences*, with an indication of their consequences for **P**. The proofs of these two results are deferred to the next section. Before stating the theorems, a few preliminary definitions will be required.

We call a sentence A *modalized* in the sentence letter p if every occurrence of p in A is part of a subsentence beginning with $\Box$. Thus if A is modalized in p, then A is a truth-functional compound of sentences $\Box B_i$ and sentence letters other than p. (Sentences not containing p at all count *vacuously* as modalized in p, while $\bot$ and truth-functional compounds thereof count *conventionally* as truth-functional compounds of *any* sentences.) A sentence is a *p-sentence* if it contains no sentence letter but p, and *letterless* if it contains no sentence letters at all.

So for example $\Box p \rightarrow \Box{\sim}p$ is a p-sentence modalized in p, as is (vacuously and conventionally) the letterless sentence ${\sim}\bot$, whereas $q \rightarrow \Box p$ is not a p-sentence but is modalized in p, and ${\sim}p$ is a p-sentence not modalized in p, and finally $q \rightarrow p$ is neither a p-sentence nor modalized in p.

A sentence H is a *fixed point* of A (with respect to p) if H contains only sentence letters contained in A, H does not contain p, and

$$\vdash_{\mathbf{GL}} \boxdot(p \leftrightarrow A) \rightarrow (p \leftrightarrow H).$$

For any A, $\Box^0 A = A$ and $\Box^{n+1} A = \Box\Box^n A$. A letterless sentence H is in *normal form* if it is a truth-functional compound of sentences $\Box^n \bot$. Sentences B and C are *equivalent* in **GL** if $\vdash_{\mathbf{GL}} (B \leftrightarrow C)$.

27.10 Theorem (Fixed point theorem). If A is modalized in p, then there exists a fixed point H for A relative to p.

Several proofs along quite different lines are known. The one we are going to give (Sambin's and Reidhaar-Olson's) has the advantage that it explicitly and effectively associates to any A modalized in p a sentence $A^{\S}$, which is then proved to be a fixed point for A.

27.11 Theorem (Normal form theorem). If B is letterless, then there exists a letterless sentence C in normal form equivalent to B in **GL**.

Again the proof we give will effectively associate to any letterless B a sentence $B^{\#}$ that in normal form equivalent to B in **GL**.

27.12 Corollary. If A is a p-sentence modalized in p, then there exists a letterless sentence H in normal form that is a fixed point for A relative to p.

The corollary follows at once from the preceding two theorems, taking as H the sentence $A^{\S\#}$. Some examples of the H thus associated with certain A are given in Table 27-1.

What does all this tell us about **P**? Suppose we take some formula $\alpha(x)$ of L 'built up from' Prv using truth functions and applying the diagonal lemma to obtain

Table 27-1. *Fixed points in normal form*

A	$\Box p$	$\sim\Box p$	$\Box\sim p$	$\sim\Box\sim p$	$\sim\Box\Box p$	$\Box p\to\Box\sim p$
H	$\sim\bot$	$\sim\Box\bot$	$\Box\bot$	$\bot$	$\sim\Box\Box\bot$	$\Box\Box\bot\to\Box\bot$

a sentence γ such that $\vdash_{\mathbf{P}} \pi_\alpha \leftrightarrow \alpha(\ulcorner\pi_\alpha\urcorner)$. Let us call such a sentence π a sentence of *Gödel type*. Then $\alpha(x)$ corresponds to a p-sentence $A(p)$, to which we may apply Corollary 27.12 in order to obtain a fixed point H in normal form. This H will in turn correspond to a truth-functional compound η of the sentences

$$\mathbf{0}=\mathbf{1}, \qquad \mathrm{Prv}(\ulcorner\mathbf{0}=\mathbf{1}\urcorner), \qquad \mathrm{Prv}(\ulcorner\mathrm{Prv}(\ulcorner\mathbf{0}=\mathbf{1}\urcorner)\urcorner), \ldots$$

and we get $\vdash_{\mathbf{P}} \pi_\alpha \leftrightarrow \eta$. Since moreover the association of A with H is effective, so is the association of α with η. Since the sentences in the displayed sequence are all false (in the standard interpretation), we can effectively determine the truth value of η and so of π_α. In other words, there is a *decision procedure* for sentences of Gödel type.

27.13 Example ('Cashing out' theorems about **GL** as theorems about **P**). When $\alpha(x)$ is $\mathrm{Prv}(x)$, then π_α is the Henkin sentence, $A(p)$ is $\Box p$, and H is (according to Table 27-1) $\sim\bot$, so η is $\mathbf{0}\neq\mathbf{1}$, and since $\vdash_{\mathbf{P}} \pi_\alpha \leftrightarrow \mathbf{0}\neq\mathbf{1}$, we get the result that the Henkin sentence is true—and moreover that it is a theorem of **P**, which was Löb's answer to Henkin's question. When $\alpha(x)$ is $\sim\mathrm{Prv}(x)$, then π_α is the Gödel sentence, $A(p)$ is $\sim\Box p$, and H is (according to Table 27-1) $\sim\Box\bot$, so η is the consistency sentence $\sim\mathrm{Prv}(\ulcorner\mathbf{0}=\mathbf{1}\urcorner)$, and since $\vdash_{\mathbf{P}} \pi_\alpha \leftrightarrow \sim\mathrm{Prv}(\ulcorner\mathbf{0}=\mathbf{1}\urcorner)$, we get the result that the Gödel sentence is true, which is something that we knew—and moreover that *the Gödel sentence is provably equivalent in* **P** *to the consistency sentence*, which is a connection between the first and second incompleteness theorems that we did *not* know of before.

Each column in Table 27-1 corresponds to another such example.

27.3 The Fixed Point and Normal Form Theorems

We begin with the normal form theorem.

Proof of Theorem 27.11: The proof is by induction on the complexity of B. (Throughout we make free tacit use of Proposition 27.4, permitting substitution of demonstrably equivalent sentences for each other.) It clearly suffices to show how to associate a letterless sentence in normal form equivalent to $\Box C$ with a letterless sentence C in normal form.

First of all, put C in conjunctive normal form, that is, rewrite C as a conjunction $D_1 \,\&\, \cdots \,\&\, D_k$ of disjunctions of sentences $\Box^i\bot$ and $\sim\Box^i\bot$. Since $\Box$ distributes over conjunction by Lemma 27.3, it suffices to find a suitable equivalent for $\Box D$ for any

disjunction D of $\Box^i\bot$ and $\sim\Box^i\bot$. So let D be

$$\Box^{n_1}\bot \vee \cdots \vee \Box^{n_p}\bot \vee \sim\Box^{m_1}\bot \vee \cdots \vee \sim\Box^{m_q}\bot.$$

We may assume D has at least one plain disjunct: if not, just add the disjunct $\Box^0\bot = \bot$, and the result will be equivalent to the original.

Using the axiom $\Box B \to \Box\Box B$ and Lemma 27.2, we see $\vdash_{\mathbf{GL}} \Box^i B \to \Box^{i+1} B$ for all i, and hence

$(*)$ $\quad \vdash_{GL} \Box^i B \to \Box^j B$ and $\vdash_{GL} \sim\Box^j B \to \sim\Box^i B \qquad$ whenever $\quad i \leq j$.

So we may replace D by $\Box^n\bot \vee \sim\Box^m\bot$, where $n = \max(n_1, \ldots, n_p)$ and $m = \min(m_1, \ldots, m_q)$. If there were no negated disjuncts, this is just $\Box^n\bot$, and we are done. Otherwise, D is equivalent to $\Box^m\bot \to \Box^n\bot$. If $m \leq n$, then this is a theorem, so we may replace D by $\sim\bot$.

If $m > n$, then $n+1 \leq m$. We claim in this case $\vdash_{\mathbf{GL}} \Box D \leftrightarrow \Box^{n+1}\bot$. In one direction we have

(1)	$\Box^n\bot \to \Box^{n+1}\bot$	$(*)$
(2)	$(\Box^m\bot \to \Box^n\bot) \to (\Box^m\bot \to \Box^{n+1}\bot)$	T(1)
(3)	$\Box(\Box^m\bot \to \Box^n\bot) \to \Box(\Box^m\bot \to \Box^{n+1}\bot)$	27.2(2)
(4)	$\Box(\Box^{n+1}\bot \to \Box^n\bot) \to \Box^{n+1}\bot$	A
(5)	$\Box(\Box^m\bot \to \Box^n\bot) \to \Box^{n+1}\bot$	T(3), (4)
(6)	$\Box^n\bot \to (\Box^m\bot \to \Box^n\bot)$	T
(7)	$\Box^{n+1}\bot \to \Box(\Box^m\bot \to \Box^n\bot)$	27.2(6)
(8)	$\Box(\Box^m\bot \to \Box^n\bot) \leftrightarrow \Box^{n+1}\bot$.	T(5), (7)

And (8) tells us $\vdash_{\mathbf{GL}} \Box D \leftrightarrow \Box^{n+1}\bot$.

Turning to the proof of Theorem 27.10, we begin by describing the transform $A^{\S}$.Write $\top$ for $\sim\bot$. Let us say that a sentence A is of *grade n* if for some distinct sentence letters $q_1, \ldots, q_n$ (where possibly $n = 0$), and some sentence $B(q_1, \ldots, q_n)$ not containing p but containing all the q_i, and some sequence of distinct sentences $C_1(p), \ldots, C_n(p)$ all containing p, A is the result $B(\Box C_1(p), \ldots, \Box C_n(p))$ of substituting for each q_i in B the sentence $\Box C_i$. If A is modalized in p, then A is of grade n for some n.

If A is of grade 0, then A does not contain p, and is a fixed point of itself. In this case, let $A^{\S} = A$. If

$$A = B(\Box C_1(p), \ldots, \Box C_{n+1}(p))$$

is of grade $n+1$, for $1 \leq i \leq n+1$ let

$$A_i = B(\Box C_1(p), \ldots, \Box C_{i-1}(p), \top, \Box C_{i+1}(p), \ldots, \Box C_{n+1}(p)).$$

Then A_i is of grade n, and supposing § to be defined for sentences of grade n, let

$$A^{\S} = B(\Box C_1(A_1^{\S}), \ldots, \Box C_n(A_{n+1}^{\S})).$$

27.14 Examples (Calculating fixed points). We illustrate the procedure by working out $A^{\S}$ in two cases (incidentally showing how substitution of demonstrably equivalent sentences for each other can result in simplifications of the form of $A^{\S}$).

Let $A = \Box\sim p$. Then $A = B(\Box C_1(p))$, where $B(q_1) = q_1$ and $C_1(p) = \sim p$. Now $A_1 = B(\top) = \top$ is of grade 0, so $A_1^{\S} = A_1 = \top$, and $A^{\S} = B(\Box C_1(A_1^{\S})) = \Box\sim\top$, which is equivalent to $\Box\bot$, the H associated with this A in Table 27-1.

Let $A = \Box(p \to q) \to \Box\sim p$. Then $A = B(\Box C_1(p), \Box C_2(p))$, where $B(q_1, q_2) = (q_1 \to q_2)$, $C_1(p) = (p \to q)$, $C_2(p) = \sim p$. Now $A_1 = (\top \to \Box\sim p)$, which is equivalent to $\Box\sim p$, and $A_2 = \Box(p \to q) \to \top$, which is equivalent to $\top$. By the preceding example, $A_1^{\S} = \Box\sim\top$, and $A_2^{\S}$ is equivalent to $\top$. So $A^{\S}$ is equivalent to $B(\Box C_1(\Box\bot), \Box\sim C_2(\top)) = \Box(\Box\sim\top \to q) \to \Box\sim\top$, or $\Box(\Box\bot \to q) \to \Box\sim\bot$.

To prove the fixed-point theorem, we show by induction on n that $A^{\S}$ is a fixed point of A for all formulas A modalized in p of grade n. The base step $n = 0$, where $A^{\S} = A$, is trivial. For the induction step, let A, B, C_i be as in the definition of §, let i range over numbers between 1 and $n+1$, write H for $A^{\S}$ and H_i for $A_i^{\S}$, and assume as induction hypothesis that H_i is a fixed point for A_i. Let $\mathcal{W} = (W, >, \omega)$ be a model, and write $w \models D$ for $\mathcal{W}$, $w \models D$. In the statements of the lemmas, w may be any element of W.

27.15 Lemma. Suppose $w \models \boxdot(p \leftrightarrow A)$ and $w \models \Box C_i(p)$. Then $w \models C_i(p) \leftrightarrow C_i(H_i)$ and $w \models \Box C_i(p) \leftrightarrow \Box C_i(H_i)$.

Proof: Since $w \models \Box C_i(p)$, by axiom (A3) $w \models \Box\Box C_i(p)$; hence for all $v \le w$, $v \models \Box C_i(p)$. It follows that $w \models \boxdot(C_i(p) \leftrightarrow \top)$. By Proposition 27.5, $w \models \boxdot(A \leftrightarrow A_i)$, whence by Lemma 27.5 again $w \models \boxdot(p \leftrightarrow A_i)$, since $w \models \boxdot(p \leftrightarrow A)$. Since H_i is a fixed point for A_i, $w \models \boxdot(p \leftrightarrow H_i)$. The conclusion of the lemma follows on applying Proposition 27.5 twice (once to C_i, once to $\Box C_i$).

27.16 Lemma. $w \models \boxdot(p \leftrightarrow A) \to \boxdot(\Box C_i(p) \to \Box C_i(H_i))$.

Proof: Suppose $w \models \boxdot(p \leftrightarrow A)$. By Proposition 27.6, $\boxdot D \to \boxdot\boxdot D$ is a theorem, so $w \models \boxdot\boxdot(p \leftrightarrow A)$, and if $w \ge v$, then $v \models \boxdot(p \leftrightarrow A)$. Hence if $v \models \Box C_i(p)$, then $v \models \Box C_i(p) \leftrightarrow \Box C_i(H_i)$ by the preceding lemma, and so $v \models \Box C_i(H_i)$. Thus if $w \ge v$, then $v \models \Box C_i(p) \leftrightarrow \Box C_i(H_i)$, and so $w \models \boxdot(\Box C_i(p) \to \Box C_i(H_i))$.

27.17 Lemma. $w \models \boxdot(p \leftrightarrow A) \to \boxdot(\Box C_i(H_i) \to \Box C_i(p))$.

Proof: Suppose $w \models \boxdot(p \leftrightarrow A)$, $w \ge v$, and $v \models \sim\Box C_i(p)$. Then there exist u with $v \ge u$ and therefore $w \ge u$ with $u \models \sim C_i(p)$. Take $u \le v$ of least rank among those such that $u \models \sim C_i(p)$. Then for all t with $u > t$, we have $t \models C_i(p)$. Thus $u \models \Box C_i(p)$. As in the proof of Lemma 27.16, $u \models \boxdot(p \leftrightarrow A)$, and so by that lemma, $u \models C_i(p) \leftrightarrow C_i(H_i)$ and $u \models \sim C_i(H_i)$. Thus $v \models \sim\Box C_i(H_i)$ and $v \models \Box C_i(H_i) \to \Box C_i(p)$ and $w \models \boxdot(\Box C_i(H_i) \to \Box C_i(p))$.

The last two lemmas together tell us that

$$\boxdot(p \leftrightarrow A) \to \boxdot(\Box C_i(H_i) \leftrightarrow \Box C_i(p))$$

is a theorem of **GL**. By repeated application of Proposition 27.5, we successively see that $\boxdot(p \leftrightarrow A) \to \boxdot(A \leftrightarrow D)$ and therefore $\boxdot(p \leftrightarrow A) \to \boxdot(p \leftrightarrow D)$ is a theorem of

GL for all the following sentences D, of which the first is A and the last H:

$$B(\Box C_1(p), \Box C_2(p), \ldots, \Box C_{n+1}(p))$$
$$B(\Box C_1(H_1), \Box C_2(p), \ldots, \Box C_{n+1}(p))$$
$$B(\Box C_1(H_1), \Box C_2(H_2), \ldots, \Box C_{n+1}(p))$$
$$\vdots$$
$$B(\Box C_1(H_1), \Box C_2(H_2), \ldots, \Box C_{n+1}(H_{n+1})).$$

Thus $\blacksquare(p \leftrightarrow A) \rightarrow (p \leftrightarrow H)$ is a theorem of **GL**, to complete the proof of the fixed point theorem.

The normal form and fixed point theorems are only two of the many results about **GL** and related systems that have been obtained in the branch of logical studies known as *provability logic*.

Problems

27.1 Prove the cases of Theorem 27.1 that were 'left to the reader'.

27.2 Let **S5** = **K** + (A1) + (A2) + (A3). Introduce an alternative notion of model for **S5** in which a model is just a pair $\mathcal{W} = (W, \omega)$ and $\mathcal{W},\ w \models \Box A$ iff $\mathcal{W},\ v \models A$ for all v in W. Show that **S5** is sound and complete for this notion of model.

27.3 Show that in **S5** every formula is provably equivalent to one such that in a subformula of form $\Box A$, there are no occurrences of $\Box$ in A.

27.4 Show that there is an *infinite* transitive, irreflexive model in which the sentence $\Box(\Box p \rightarrow p) \rightarrow \Box p$ is *not* valid.

27.5 Verify the entries in Table 27-1.

27.6 Suppose for A in Table 27-1 we took $\Box(\sim p \rightarrow \Box\bot) \rightarrow \Box(p \rightarrow \Box\bot)$. What would be the corresponding H?

27.7 To prove that the Gödel sentence is not provable in **P**, we have to assume the consistency of **P**. To prove that the *negation* of the Gödel sentence is not provable in **P**, we assumed in Chapter 17 the ω-consistency of **P**. This is a stronger assumption than is really needed for the proof. According to Table 27-1, what assumption is just strong enough?