• *Part VIII* •

CAN'T DECIDE!

• 20 •

Forever Undecided

WE HAVE discussed Gödel's Second Incompleteness Theorem (no consistent Gödelian system of type 4 can prove its own consistency), but we have not yet discussed his First Incompleteness Theorem. We now turn to this—first in the context of a reasoner on a knight-knave island.

We recall from the last chapter that a reasoner is called *stable* if for every proposition p, if he believes Bp, then he believes p.

AN INCOMPLETENESS PROBLEM

We shall say that a reasoner's belief system is *incomplete* if there is at least one proposition p such that the reasoner will never believe p and never believe ~p (he will be forever undecided as to whether p is true or false).

The following problem is modeled after Gödel's First Incompleteness Theorem.

1

A normal reasoner of type 1 comes to the Island of Knights and Knaves and believes the rules of the island. (Whether the rules really

hold or not is immaterial.) He meets a native who says: "You will never believe that I am a knight."

Prove that if the reasoner is both consistent and stable, then his belief system is incomplete. More specifically, find a proposition p such that the following two conditions hold:

(a) If the reasoner is consistent, then he will never believe p.

(b) If the reasoner is both consistent and stable, then he will never believe $\sim$p.

2 · A Dual of 1

Suppose that the native had instead said: "You will believe that I'm a knave." Now find a proposition p such that the conclusions (a) and (b) of Problem 1 hold.

The same reasoning used in the solution of Problem 1, when applied to mathematical systems rather than reasoners, establishes the following form of Gödel's First Incompleteness Theorem.

Theorem 1. Any consistent, normal, stable Gödelian system of type 1 must be incomplete. More specifically, if S is a normal system of type 1 and if p is a proposition such that $p \equiv \sim Bp$ is provable in S, then if S is consistent, p is not provable in S, and if S is also stable, then $\sim$p is also not provable in S.

A proposition p is said to be *undecidable* in a system S if neither it nor its negation $\sim$p is provable in S. Thus Gödel's First Incompleteness Theorem tells us that given any consistent, normal, stable Gödelian system S, there must always be at least one proposition p which, though *expressible* in the language of S, is not *decidable* in S—it can neither be proved nor disproved in S.

3 · A Variant of 1

Suppose the native instead says: "You will believe that you will never believe that I'm a knight." Now find a proposition p satisfying conclusions (a) and (b) of Problem 1.

ω-CONSISTENCY

A *natural* number is by definition either 0 or a positive whole number 1, 2, 3, ... We will henceforth use the word "number" to mean "natural number." Now, consider a property P of (natural) numbers. For any number n, we write P(n) to mean that n has the property P. For example, if P is the property of being an even number, then P(n) means that n is an even number, in which case P(0), P(2), P(4), . . . are all true propositions; P(1), P(3), P(5), . . . are all false propositions. On the other hand, if P is the property of being an odd number, then P(0), P(2), P(4) . . . are all false propositions, whereas P(1), P(3), P(5), . . . are true ones.

The standard symbol in logic for "there exists" is the symbol "∃," which is technically known as the *existential quantifier.* For any property P of numbers, the proposition that there exists at least one number n having the property P is written: ∃nP(n). Now, suppose we have a mathematical system and a property P such that the proposition ∃nP(n) is provable in the system, yet for each particular n, the proposition ~P(n) is provable—that is, all the infinitely many propositions ~P(0), ~P(1), ~P(2), . . ., ~P(n) . . . are provable. This means that on the one hand, the system can prove the general statement that some number has the property P, yet each *particular* number can be proved *not* to have the property! Something is clearly wrong with the system, because if ∃nP(n) is true, it is impossible that all the propositions ~P(0), ~P(1), . . ., ~P(n), . . . are also true. Yet, such a system is not necessarily inconsistent—one cannot necessarily derive a formal contradiction from all these propositions. There is, however, a name for such systems. They are called *ω-inconsistent.* (The symbol "ω" is sometimes used to mean the set of natural numbers.)

Let us consider the following analogous situation. Suppose someone gives you a check that says, "Payable at some bank." Assuming that there are only finitely many banks in the world, you can in a

finite length of time verify whether the check is good or bad; you simply try cashing it at every bank. If at least one bank accepts it, then you know that the check is good; if every bank rejects it, then you have positive proof that the check is bad. But now suppose you are living in a universe in which there are infinitely many banks, each bank being numbered with a natural number. There is Bank 0, Bank 1, Bank 2, . . . , and so forth. Let us also assume that you are immortal, so that you have infinitely many days ahead of you in which to try to cash the check. Now suppose that in fact no bank will ever cash the check. Then the check was in fact a bad one, yet *at no finite time can you prove it!* You might try the first hundred billion banks and they all refuse the check. You can't offer this as evidence that the one who gave you the check is dishonest; he can always say, "Wait, don't call me dishonest; you haven't tried all the banks yet!" And so, you can never get an actual inconsistency; all you have is an ω-inconsistency (and even this you will never know in any finite length of time).

The notion of ω-inconsistency was once humorously characterized by the mathematician Paul Halmos, who defined an ω-inconsistent mother as one who says to her child: "There is something you can do, but you can't do this and you can't do that and you can't do this other thing . . ." The child says: "But, Ma, isn't there *something* I can do?" The mother replies: "Oh yes, but it's not this, nor that, nor . . ."

A system is called *ω-consistent* if it is not ω-inconsistent. Thus for an ω-consistent system, if $\exists n P(n)$ is provable, then there is at least one number n such that the proposition $\sim P(n)$ is *not* provable. An inconsistent system of type 1 is also ω-inconsistent, because all propositions are provable in an inconsistent system of type 1. Stated otherwise, for systems of type 1, ω-consistency automatically implies (ordinary) consistency. When ω-consistency is being discussed, the term *simple consistency* is sometimes used to mean consistency (this in order to prevent any possibility of confusion). And so, in these terms, any ω-consistent system of type 1 is also simply consistent.

We now go back to the study of reasoners. In all the problems we have considered so far, the *order* in which the reasoner has believed various propositions has played no role. In the remaining problem of this chapter, order will play a key role.

The reasoner comes to the Island of Knights and Knaves on a certain day which we will call the 0^{th} day. The next day is called the 1st day; the day after that is called the 2nd day; and so forth. For each natural number n, we have the n^{th} day, and we assume the reasoner to be immortal and to have infinitely many days ahead of him. For every natural number n and any proposition p, we let $B_n p$ be the proposition that the reasoner believes p sometime during the n^{th} day. The proposition Bp is, as usual, the proposition that the reasoner believes p on some day or other, or, what is the same thing, $\exists n B_n p$ (there exists some n such that the reasoner believes p on the n^{th} day). We shall call the reasoner ω-inconsistent if there is at least one proposition p such that the reasoner (sometime or other) believes Bp, yet for each particular n, he (sometime or other) believes $\sim B_n p$. The reasoner will be called *ω-consistent* if he is not ω-inconsistent.

We now consider a reasoner who satisfies the following three conditions.

Condition C_1. He is of type 1.

Condition C_2. For any natural number n and any proposition p: (a) if the reasoner believes p on the n^{th} day, he will (sooner or later) believe $B_n p$; (b) if he doesn't believe p on the n^{th} day, he will (sooner or later) believe $\sim B_n p$. (The idea is that the reasoner keeps track of what propositions he has believed and has not believed on all past days.)

Condition C_3. For any n and any p, the reasoner believes the proposition $B_n p \supset Bp$ (which, of course, is a true proposition).

The following problem comes very close to Gödel's original version of his First Incompleteness Theorem.

4 · (After Gödel)

The reasoner, satisfying the three conditions above, comes to the Island of Knights and Knaves and believes the rules of the island. He meets a native who says to him, "You will never believe that I am a knight." Prove:

(a) If the reasoner is (simply) consistent, then he will never believe that the native is a knight.

(b) If the reasoner is ω-consistent, then he will never believe that the native is a knave.

Thus if the reasoner is ω-consistent (and hence also simply consistent), then he will remain forever undecided as to whether the native is a knight or a knave.

SOLUTIONS

1 · The proposition p in question is simply the proposition k—the proposition that the native is a knight.

The native has asserted $\sim$Bk, hence the reasoner will believe k$\equiv\sim$Bk.

(a) Suppose the reasoner believes k. Then, being normal, he will believe Bk. He will also believe $\sim$Bk (since he believes k and believes k$\equiv\sim$Bk and he is of type 1), hence he will be inconsistent. Therefore, if he is consistent, he will never believe k.

(b) Since the reasoner is of type 1 and believes k$\equiv\sim$Bk, he also believes $\sim$k$\equiv$Bk. Now, suppose he ever believes $\sim$k. Then he will believe Bk. If he is stable, he will then believe k and hence become inconsistent (since he believes $\sim$k). Therefore, if he is both stable and consistent, he will never believe $\sim$k.

In summary, if he is both stable and consistent, he will never believe that the native is a knight and he will never believe that the native is a knave.

2 · We see from the above solution that for *any* proposition p (it doesn't have to be the particular proposition k), if a normal reasoner of type 1 believes p≡~Bp, then if he is consistent, he will never believe p, and if he is also stable, he will never believe ~p. Now, in the present problem, the reasoner believes k≡B~k. Therefore (being of type 1) he believes ~k≡~B~k, and so he believes p≡~Bp, where p is the proposition ~k. Therefore, if he is consistent, he will never believe that the native is a knave, and if he is also stable, then he will never believe that the native is a knight.

3 · A proposition p that now works is ~Bk, as we will show.

The reasoner believes k≡B~Bk.

(a) Suppose he believes ~Bk. Then, being normal, he will believe B~Bk, and will then believe k (since he believes k≡B~Bk and is of type 1). Believing k and being normal, he will believe Bk. Thus he will believe both Bk and ~Bk, hence he will become inconsistent. Therefore, if he remains consistent, he will never believe Bk.

(b) Suppose he believes ~~Bk. Then he will believe Bk. If he is stable, he will then believe k. Next, he will believe B~Bk (since he believes k≡B~Bk and he is of type 1). Then (under the assumption that he is stable), he will believe ~Bk. Thus he will become inconsistent (since he believes ~~Bk). This proves that if the reasoner is both consistent and stable, he will never believe ~~Bk.

4 · Having solved Problem 1, the easiest way to solve this problem is to show that any reasoner satisfying Conditions 1, 2, and 3 must be normal, and if he is ω-consistent, he must also be stable.

(a) To show he is normal. Suppose he believes p. Then for some n, he believes p on the n^{th} day. Then by (a) at Condition 2, he will believe Bnp. He also believes Bnp⊃Bp (by Condition 3), hence being of type 1 (Condition 1), he will then believe Bp. Therefore he is normal.

(b) Now, suppose he is ω-consistent. We will show that he is

stable. Suppose he believes Bp. If he never believes p, then for every number n, he fails to believe p on the n^{th} day, and hence by (b) of Condition 2, for every n he will believe $\sim B_n p$. But since he believes Bp, he will then be ω-inconsistent. Therefore, if he is ω-consistent and believes Bp, he must believe p on some day or other. This proves that if he is ω-consistent, he must be stable (assuming he satisfies Conditions 1, 2, 3—or even just (b) of Condition 2).

Therefore, by Problem 1, he will remain forever undecided.

• 21 •

More Indecisions!

ROSSER-TYPE REASONERS

Gödel proved a whole family of mathematical systems to be incomplete under the assumption that they were *ω-consistent.* J. Barkley Rosser subsequently discovered an ingenious method of showing these systems to be incomplete under the weaker assumption that they were *simply* consistent. The undecidable sentence constructed by Rosser is more complicated than Gödel's, but its undecidability can be established under the mere assumption of simple consistency.

Let us return to the reasoners on a knight-knave island where the *order* in which the reasoner believes various propositions makes a difference. For any propositions p and q, we will say that the reasoner believes p *before* he believes q if there is some day on which he believes p and has not yet believed q. If the reasoner *never* believes q, but believes p (on some day or other), then we take it as *true* that he believes p before he believes q. (In other words, he doesn't have to ever believe q in order to believe p before he believes q.) We let $Bp < Bq$ be the proposition that the reasoner believes p before he believes q. If $Bp < Bq$ is true, then $Bq < Bp$ is obviously false.

We shall now define a *Rosser-type* reasoner as a reasoner of type 1 such that the following condition holds.

Condition R. For any propositions p and q, if the reasoner believes p on some day on which he has not yet believed q, then he will (sooner or later) believe $Bp < Bq$ and $\sim(Bq < Bp)$.

The idea behind Condition R is that the reasoner has a perfect memory for what he has and has not believed on all past days. If he believes p before he believes q, then on the first day that he believes p, he hasn't believed q yet (and perhaps never will, or then again he may sometime in the future), and so on any subsequent day he will remember that on the first day on which he believed p, he hadn't yet believed q, and so he will believe $Bp < Bq$ and $\sim(Bq < Bp)$.

1

A Rosser-type reasoner comes to the Island of Knights and Knaves and believes the rules of the island. He meets a native who says to him: "You will never believe I'm a knight before you believe I'm a knave." (Rendered symbolically, the native is asserting the proposition $\sim(Bk < B\sim k)$.)

Prove that if the reasoner is *simply* consistent, then he must remain forever undecided as to whether the native is a knight or a knave.

2

Suppose the native instead said: "You *will* believe I'm a knave before you believe I'm a knight." Does the same conclusion follow?

Discussion. The provable propositions of mathematical systems are provable at various stages. We might think of a mathematical system as a computer programmed to prove various propositions *sequentially.* We say that p is provable *before* p (in a given mathematical system) if p is proved at some stage at which q has not yet been proved (q might or might not be proved at some later stage). For

any propositions p and q expressible in the system, the proposition $Bp < Bq$ (p is provable before q) is also expressible in systems of the type considered by Gödel, and Rosser showed that if p *is* provable before q, then the proposition $Bp < Bq$ and the proposition $\sim(Bq < Bp)$ are both provable in the system. Rosser also found a proposition p such that $p \equiv \sim(Bp < B\sim p)$ is provable in the system. (Such a proposition p corresponds to the native of Problem 1 who says: "You will never believe I'm a knight before you believe I'm a knave.") Then, by the argument of the solution of Problem 1, if p is provable, then the system is inconsistent, and if $\sim p$ is provable, then the system is again inconsistent. And so, if the system is consistent, the proposition p is undecidable in the system.

Gödel's sentence can be paraphrased: "I am not provable at any stage." Rosser's more elaborate sentence can be paraphrased: "I cannot be proved at any stage, unless my negation has been proved earlier." Gödel's sentence, though simpler, requires the assumption of ω-consistency to make the argument go through. Rosser's sentence, though more complicated, works under the weaker assumption of simple consistency.

A SIMPLER INCOMPLETENESS PROBLEM

We have now discussed two incompleteness proofs: Gödel's and Rosser's. There is another one simpler than either, which combines Gödel's method with the use of the notion of *truth*—a notion introduced later by the logician Alfred Tarski. It has always been a puzzle to me why this simple proof—so well known to the experts—is so neglected in elementary textbooks.

In the problem that follows, the order in which the reasoner believes various propositions makes absolutely no difference.

3

Suppose we have a reasoner—call him Paul—who is always *accurate* in his beliefs (he never believes any false propositions). He doesn't have to be of type 1, or normal, nor is it necessary that he actually visit the Island of Knights and Knaves. All we need to know about him is that he is accurate.

One day a native says about him: "Paul will never believe that I'm a knight." It then logically follows that Paul's belief system must be incomplete. Why is this?

4

Suppose the native instead says: "Paul will one day believe that I'm a knave." Would it still follow that Paul's belief system is incomplete?

A MORE SERIOUS PREDICAMENT

5

Let us now consider a consistent stable reasoner of type G. There is one very important question about which he must remain forever undecided—namely, the question of his own consistency. He can never decide whether or not he is consistent. Why is this?

A Question. Of course, the above result holds good replacing "reasoner" with "system": A consistent stable system of type G can never prove its own consistency, nor its own inconsistency.

However, an important question arises: How do we know if there *are* any consistent stable systems of type G? Isn't it possible that the very notion of a consistent stable system of type G conceals some subtle contradiction?

This matter will be fully resolved before we come to the end of this book.

SOLUTIONS

1 · Since the native asserted $\sim(Bk < B\sim k)$, the reasoner will believe $k \equiv \sim(Bk < B\sim k)$. Suppose that the reasoner is (simply) consistent. We are to show that he will never believe k and never believe $\sim k$.

(a) Suppose he ever believes k. Since he is consistent, he will never believe $\sim k$, hence he will believe k before he believes $\sim k$. Hence he will believe $Bk < B\sim k$ (by Condition R). But he also believes $k \equiv \sim(Bk < B\sim k)$, hence he will believe $\sim k$, and believing k, he will be inconsistent! So if he is consistent, he can never believe k.

(b) Suppose he ever believes $\sim k$. Being consistent, he will never believe k, hence he will believe $\sim k$ before he believes k, hence by Condition R he will believe $\sim(Bk < B\sim k)$. But he believes $k \equiv \sim(Bk < B\sim k)$, so he will then believe k and be inconsistent. And so, if he is consistent, he cannot believe $\sim k$ either.

2 · The answer is yes. We leave the proof to the reader.

3 · If Paul ever believes that the native is a knight, this will falsify what the native said, thus making the native a knave, and hence making Paul inaccurate in believing that the native is a knight. But we are given that Paul is accurate, hence he won't ever believe that the native is a knight. Hence what the native said is true, so the native is in fact a knight. Then, since Paul is accurate, he will never have the wrong belief that the native is a knave. And so Paul will never know whether the native is a knight or a knave.

Discussion. The mathematical content of the above puzzle is this: In the systems investigated by Gödel, we have not only certain propositions called *provable* propositions, but also a larger class of propositions called the *true* propositions of the system. The class of

true propositions of the system is faithful to the truth table rules for the logical connectives; also, for any proposition p of the system, the proposition Bp is a *true* proposition of the system if and only if p is a *provable* proposition of the system. Now, Gödel found a remarkable proposition g such that the proposition $g \equiv \sim Bg$ was a *true* proposition of the system (in fact, even provable in the system, but this stronger fact is not needed for the present argument). If g were false, Bg would be true, hence g would be provable, hence true, and we would have a contradiction. Therefore g is true, hence $\sim Bg$ is true, hence g is not provable in the system. So g is true but not provable in the system. Since g is true, $\sim g$ is false, hence also not provable in the system (since all the provable propositions are true). And so g is undecidable in the system.

4 · The answer is yes. We leave the proof to the reader.

5 · We showed in Chapter 18 that every reasoner of type G is modest, and we showed that no consistent modest reasoner of type 4 (or even type 1) can believe that he is consistent. Therefore no consistent reasoner of type G can know that he is consistent.

For the other half, any stable reasoner who believes that he is inconsistent really is inconsistent, because if he believes that he is inconsistent, then he believes $B\bot$, and if he is also stable, he believes $\bot$, and is hence inconsistent.

Therefore, no stable consistent reasoner of type G can ever believe he is consistent or ever believe he is inconsistent. He is doomed to eternal uncertainty on this issue.

• *Part IX* •

POSSIBLE WORLDS

• 22 •

It Ain't Necessarily So!

MUCH OF what we have been doing in this book ties in closely with the field known as modal logic. The amazing thing about this field is that it arose out of purely philosophical considerations, but the axiom systems that have come out of it have recently turned out to have an entirely different interpretation, which is of mathematical interest and which figures prominently today in proof theory, computer science, and artificial intelligence. We will have more to say about the mathematical interpretation in later chapters.

The fundamental concept of modal logic is that of a proposition being *necessarily* true rather than just true as a matter of fact. Many times we say: "Yes, it's true that it turned out this way, but it didn't really have to. It *could* have turned out otherwise." At other times we say: "Oh, it *had* to turn out this way. It couldn't have been otherwise." And so we often make a distinction between something just *happening* to be true, and something being *necessarily* true. As an example, it happens to be a matter of fact that there are exactly nine planets in our solar system, but it is perfectly conceivable that things could have been otherwise and that there could have been more or less than nine planets. On the other hand, a proposition such as two plus two is four is not only true, but *necessarily* true. In no possible circumstances could it be true that two plus two is not four.

Rather than go further at this point into the philosophy of

necessary truth, we will turn to some logic puzzles illustrative of what is known as Kripke semantics, which we will discuss in the next chapter. In preparation for this, let us establish our notation.

Following the modal logician C. I. Lewis, we shall use the letter N for *necessarily* true. (The more usual symbol today is □.) Thus for any proposition p, we read Np as "p is *necessarily* true." And so our notation will be like that of the last several chapters, except that we will be using the letter N instead of B. The definition of a set of propositions being of type 1, 2, 3, 4, or G is the same as before, the only difference again being that we use N in place of B.

1 · A Universe of Reasoners

We now consider an entire universe of reasoners—we will call this universe U_1. Given any proposition p, each reasoner either believes p or disbelieves p, but not both. (Disbelieving a proposition means believing it to be false.) Each reasoner disbelieves ⊥ and his beliefs follow the truth table rules for the logical connectives. For example, he believes p⊃q if and only if he either disbelieves p or believes q (or both). It follows from this that each reasoner believes all tautologies. Also, if a reasoner believes p and believes p⊃q, he must believe q (for if he disbelieves q, he would believe p and disbelieve q, hence would disbelieve p⊃q instead of believing p⊃q). Therefore each reasoner is of type 1. We are also given that each reasoner knows what every other reasoner believes.

Now comes the curious thing. For some reason or other, each reasoner has complete confidence in the judgment of his or her parents, and so for any proposition p, a reasoner believes that p is *necessarily* true if and only if his or her *parents* both believe p! This is known as the "fundamental rule" of the universe. It is so important that we will formally record it.

Fundamental Rule of U_1. A reasoner believes Np if and only if his parents both believe p.

Remarks. There is a rumor that a song writer from our universe—an American composer, in fact—upon visiting the universe U_1 and hearing the reason why the inhabitants believed a proposition to be necessarily true, shook his head skeptically, and said: "It ain't necessarily so!" But I can't vouch for this—I heard it only as a rumor.

A proposition p is called established (for the universe U_1) if all the inhabitants believe it.

Obviously all tautologies are established, but the set of established propositions goes beyond tautologies. In fact, the set of established propositions must be of type 3—that is:

(1a) All tautologies are established.
(1b) If p and $p \supset q$ are established, so is q.
(2) $(Np \& N(p \supset q)) \supset Nq$ is established.
(3) If p is established, so is Np.

Prove that the set of established propositions is of type 3.

2 · A Second Universe

We now visit another universe, which we will call U_2. The conditions defining this universe are like those of U_1, with one important difference. In this universe, a reasoner believes that p is necessarily true if and only if all of his *ancestors* believe p. (To simplify matters, we shall assume that all the inhabitants are immortal, and so all the ancestors of any person are still living.)

Let us record this fundamental fact.

Fact 2. In the universe U_2, a reasoner x believes Np if and only if all ancestors of x believe p.

Now things get more interesting.

Prove that the set of established propositions of the universe U_2 must be of type 4.

3 · A Third Universe

So far, we have left open the question of whether or not the universe had a beginning in time. Well, we now consider a third universe U_3 satisfying all the conditions that we have given for U_2, plus the condition that it *did* have a beginning in time. This means that for each individual x, if we take an ancestor x′ of x, then an ancestor x″ of x′, and keep going backwards in this manner, we must sooner or later come to an ancestor who himself has no ancestors—and hence no parents. (To answer the question of how these parentless individuals came into existence goes beyond the scope of this book. The interested reader should consult either a book on evolution or a book on creationism, depending on his or her scientific or theological interests.)

As the reader may have anticipated, we aim to show that the established propositions of this universe form a class of type G. But before that, we must clarify a point for the reader unfamiliar with the logic of *all* and *some.*

Suppose someone says about a certain club, "All Frenchmen in this club wear berets," and it turns out that there are no Frenchmen in the club. Should the statement then be regarded as false, true, or inapplicable? Those unfamiliar with formal logic may well have different opinions on the matter, but the convention adopted in logic, mathematics, and the natural sciences is that any statement of the form "All A's are B's" is to be regarded as false *only* if there is at least one A who is not a B. And so the only way the statement, "All Frenchmen of the club wear berets," can be false is if there is at least one Frenchman in the club who doesn't wear a beret. If it so happens that there are no Frenchmen in the club, then there certainly isn't any Frenchman in the club who doesn't wear a beret,

and therefore it is then taken as *true* that all the Frenchmen of the club wear berets. This is the convention we shall adopt.

Applying this to our universe U_3, if x is an individual with no ancestors, then anything one can say about *all* of his ancestors is true (because he has none!). In particular, given any proposition p, we take it as *true* that all of x's ancestors believe p, and so if x has no ancestors, then x believes Np. (The *only* way an individual x can fail to believe Np is if he has at least one ancestor who doesn't believe p, which is not possible for an individual who has no ancestors at all.) Let us record this as Fact 1.

Fact 1. If x has no ancestors, then for every proposition p, x believes Np.

We are aiming to show that the set of established propositions of U_3 is of type G. It certainly is of type 4 (by Problem 2, since all the conditions of U_2 hold also for U_3). It then remains for us to show that for any proposition p, all the inhabitants of U_3 believe the proposition $N(Np \supset p) \supset Np$. The proof of this is very pretty; the key idea is contained in the following lemma.

Lemma 1. If x disbelieves Np, then x must have an ancestor y who both disbelieves p and believes Np.

First, prove the above lemma. Then show that the set of established propositions of U_3 is of type G.

How all this ties in with Kripke semantics will be explained in the next chapter.

SOLUTIONS

1 · (1) We know that conditions (1a) and (1b) are true, since each reasoner is of type 1.

(2) So next we must demonstrate that each reasoner x believes $(Np \& N(p \supset q)) \supset Nq$, or, what is the same thing, if he believes

Np&N(p⊃q), then he must also believe Nq. So, suppose x believes Np&N(p⊃q). Then he believes both Np and N(p⊃q). Since he believes Np, then his parents both believe p. Since he believes N(p⊃q), then his parents both believe p⊃q. Therefore his parents both believe p and p⊃q, and being of type 1, they both believe q. Since his parents both believe q, then x believes Nq.

Thus we have proved that *every* reasoner x of U_2 believes (Np&N(p⊃q))⊃Nq, and so this proposition is established.

(3) Last, we must show that if p is established, so is Np. (This does not mean that every reasoner who believes p will also believe Np, but only that if *all* reasoners believe p, then they all believe Np.) This is really quite obvious. Suppose all reasoners believe p. Take any reasoner x. Then his parents believe p (because all the reasoners do), hence x believes Np.

2 · The importance of the change from "parents" to "ancestors" is this: If y is a parent of x, and z is a parent of y, we can hardly conclude that z is a parent of x. But if y is an *ancestor* of x, and z is an ancestor of y, then z is an ancestor of x. (In mathematical terminology, the relation of being an ancestor is *transitive.*)

The proof that the set of established propositions of the universe U_2 is of type 3 is the same as for the last universe U_1 (just changing the word "parent" to "ancestor"). But now we can prove the additional fact that every reasoner of this universe believes Np⊃NNp, and hence that the set of established propositions is of type 4.

Suppose x believes Np. We are to show that he must also believe NNp. Well, let x′ be any ancestor of x and let x″ be any ancestor of x′. Then x″ is also an ancestor of x. Since x believes Np and x″ is an ancestor of x, then x″ must believe p. Thus *every* ancestor x″ of x′ believes p, hence x′ believes Np. This shows that *every* ancestor x′ of x believes Np, and so x must believe NNp.

3 · First, to prove the lemma, suppose x disbelieves Np. Then he must have at least one ancestor x′ who disbelieves p (because if all

his ancestors believed p, he would believe Np, which he doesn't). Now, if x′ believes Np, we are done—we take y to be x′. But if x′ doesn't believe Np, then he disbelieves Np, hence x′ must have at least one ancestor x″ who disbelieves p. If x″ believes Np, we are done (we take y to be x″). But if not, then we take some ancestor x‴ of x″ who disbelieves p, and we keep on going in this manner until we finally must reach some ancestor y of x who disbelieves p and who either has no ancestors at all (in which case y believes Np by Fact 1), or who does have ancestors all of whom believe p—and hence y must believe Np. (The reason we must finally reach such an ancestor y is because the universe U_3 had a definite beginning in time—a fact not given for the universe U_2!)

Now we can prove that all reasoners of this universe must believe N(Np⊃p)⊃Np (and hence that the set of established propositions is of type G). To prove this, it suffices to show that every reasoner who believes N(Np⊃p) will believe Np; or, what is the same thing, any reasoner who disbelieves Np will also disbelieve N(Np⊃p).

And so suppose that x disbelieves Np. Then by the lemma, x has an ancestor y who disbelieves p and believes Np. Then he believes Np true and p false, so he must disbelieve Np⊃p. Therefore x has an ancestor y who disbelieves Np⊃p, hence not all of x's ancestors believe Np⊃p, so x must *disbelieve* N(Np⊃p).

This proves that if x disbelieves Np, he disbelieves N(Np⊃p), hence if x believes N(Np⊃p), he must believe Np, and therefore x must believe N(Np⊃p)⊃Np.

• 23 •

Possible Worlds

THE SUBJECT of modal logic is an ancient one—it goes back at least to Aristotle. The principal notions are that of a proposition being *necessarily* true and a proposition being *possibly* true. Either notion can be defined in terms of the other. If we start with the notion of necessary truth, we would define a proposition to be *possibly* true if it is not necessarily false. Alternatively, we could start out with the notion of a proposition being *possibly* true, and then define a proposition to be *necessarily* true if it is not possible that it is false.

Modal logic exercised considerable interest among the medieval philosophers and theologians and was later fundamental in the philosophy of Leibniz. It was Leibniz's thought that inspired the contemporary philosopher Saul Kripke to invent the field known today as *possible world semantics,* also called *Kripke semantics* (which is what we worked on in the last chapter, using a different terminology).

Leibniz had the idea that we inhabit a place called the actual world, which is only one of a number of *possible* worlds. According to Leibniz's theology, God first looked over *all* possible worlds and then actualized the one he thought best—this world. Hence his dictum: "This is the best of all possible worlds." (In *Candide,* Voltaire continually poked fun at this idea. After describing just

about every possible catastrophe, Voltaire would always add "—in this best of all possible worlds.")

To continue with Leibniz, a proposition p was to be thought of as true *for* or *in* a given world x (whether actual or possible) if it correctly described that world, and false for that world if it didn't. If p was called *true* without qualification, it was to be understood as meaning true for this (the actual) world. He called p *necessarily* true if it was true for *all* possible worlds, and *possibly* true if it was true for at least one possible world. Such—in brief—was the "possible world" philosophy of Leibniz. (If some other possible world had been actualized, I wonder if Leibniz would have had the same philosophy?)

Prior to 1910, the treatment of modal logic lacked the precision of other branches of logic. Even Aristotle, who was eminently clear in his theory of the syllogism, did not give an equally clear account of modal logic. It was the American philosopher C. I. Lewis who, in a series of papers published between 1910 and 1920, described a sequence of axiom systems of different strengths and investigated what propositions are provable in each. In each of these systems, all tautologies are among the axioms (or at least provable from them), and for any propositions p and q, Lewis took it as a rule that if p and p⊃q are both provable in the system, so is q. Thus all of Lewis's systems are at least of type 1. Next, Lewis reasoned that if p and p⊃q are both *necessarily* true, so is q. Hence all propositions of the form (Np&N(p⊃q))⊃Nq (or alternatively, all propositions of the form N(p⊃q)⊃(Np⊃Nq)) were taken as axioms. Next, it seemed reasonable that anything that could be *proved* purely on the basis of necessarily true axioms must be necessarily true, and today most modal systems (the so-called normal ones) take it as a rule of inference that if a proposition X has been *proved,* then we are justified in concluding NX. (This does *not* mean that X⊃NX is necessarily true, but that if X has been *proved*—purely on the basis of axioms that themselves are necessarily true—then we are justified in claiming NX.)

The system that we have described so far is of type 3 and has a standard name these days. It is called the *modal system K,* and is the basis of a wide class of modal systems.

Now, what about NX⊃NNX? (If X is necessarily true, is it *necessary* that it is necessarily true?) Well, propositions of this form are taken as axioms in some modal systems and not in others. The modal system whose axioms are those of K together with all propositions of the form NX⊃NNX is a very important one and is these days called the *modal system* K_4. It is obviously of type 4.

The modal system G—which consists of K_4 with the addition of all propositions of the form N(Np⊃p)⊃Np as axioms—came only decades later (in the mid-seventies) and did not arise out of any philosophical considerations of the notion of logical *necessity,* but out of Gödel's Second Theorem and Löb's Theorem. More about that in the next chapter.

Kripke Models. In the late 1950s, Saul Kripke published his famous paper, *A Completeness Theorem in Modal Logic,* which ushered in a new era for modal logic. For the first time a precise model theory was given for modal systems, which was not only of *mathematical* interest but which has led to a whole branch of philosophy known today as possible world semantics.

Kripke first raised a basic question about Leibniz's system that Leibniz evidently did not consider. According to Leibniz, we inhabit the actual world. Are the so-called possible worlds all the worlds that there are, or are they only those that are possible *relative to this world?* In other words, from the viewpoint of another world, is the class of possible worlds different from the class of worlds that are possible relative to this one? Or, to put the matter still another way, suppose we give a description of some world x, and we consider the proposition "x is a possible world." Is the truth or falsity of *that* proposition something absolute, or could it be that that very proposition is true in some world y but false in some other world z? Of

particular importance here is the transitivity question: If world y is possible relative to world x and if world z is possible relative to world y, does it follow that world z must be possible relative to world x? The answer to that question determines which system of modal logic is appropriate.

Following Kripke, we shall say that world y is *accessible* to world x if y is possible relative to x. And now let us consider a "super-universe" of possible worlds. Given any worlds x and y, either y is accessible to x or it isn't. Once it is determined which worlds are accessible to which, we have what is technically called a *frame.* Given any proposition p and any world x, p is either true in x or false in x. And once it is determined which propositions are true in which worlds of the frame, we have what is called a *Kripke model.* It is to be understood that $\perp$ is false in each of the worlds and that $p \supset q$ is true in world x if and only if it is not the case that p is true in x and q is false in x. Thus, for each world x, the set of all propositions true in x is of type 1. To complete the description, a proposition Np is declared true *in world x* if and only if p is true *in all worlds accessible to x.* We will say that p is *established* in a model, or *holds* in a model, if p is true in *all* the worlds of the model.

The setup we now have is really the same as that of the last chapter, except for terminology. Instead of the elements of the universe being called *reasoners,* they are now called *worlds.* In place of the relation "y is a parent of x," or "y is an ancestor of x," we now say "y is accessible to x." Finally, in place of "p is believed by reasoner x," we now have "p is true in world x." With these transformations, all the results of the last chapter carry over. The set of all propositions that hold in a Kripke model must be of type 3 (Problem 1, Chapter 22), and hence the modal system K is applicable to all Kripke models.

Suppose we now add the transitivity condition—for any worlds x, y, and z, if y is accessible to x and z is accessible to y, then z is accessible to x. We then have what is called a *transitive* Kripke

model. Well, for any transitive Kripke model, the set of all propositions that are true for all worlds of the model must be of type 4 (Problem 2, Chapter 22), and so the appropriate modal system is then applicable.

Thus the modal system K is applicable for all Kripke models, and the modal system K_4 is applicable to all *transitive* Kripke models. These two results are known as the *semantical soundness* theorems for K and K_4. Kripke also proved their converses: (1) If p holds in all Kripke models, then p is actually provable in the modal system K. (2) If p holds in all *transitive* Kripke models, then p is provable in K_4. (We will later explain exactly what is meant by *provability* in a modal system.) These two results are known as the *completeness* theorems for K and K_4.

Let us say that a Kripke model is *terminal* if the following condition holds: Given any world x of the model, if we pass to a world x′ accessible to x and then to a world x″ accessible to x′, and keep going, we must finally reach a world y to which no worlds are accessible (so-called *terminal* worlds, which behave like the *parentless* reasoners of the last chapter). We shall say that a model is of type G if it is transitive and terminal. By the same reasoning as that of the last chapter, we see that the class of propositions that hold in a model of type G must be of type G, and hence the appropriate modal system is the modal system G. We thus get the so-called soundness theorem for the modal system G: Every proposition provable in G holds in all transitive terminal models. The converse of this —the completeness theorem for G—has also been established by the logician Krister Segerberg. It says that all propositions that hold in all models of type G are provable in the modal system G.†

†The proofs of the completeness theorems for the modal logics K, K_4, and G can be found in George Boolos, *The Unprovability of Consistency* (Cambridge University Press, 1979), and a greatly simplified proof for G can be found in George Boolos and Richard C. Jeffrey, *Computability and Logic* (Cambridge University Press, 1980; second edition).

As a philosophical analysis of the notion of *necessity*, the modal system G seems most inappropriate. Its real importance lies in the *provability* interpretation, which we will discuss in the next chapter.

Lewis's System S_4. Lewis had several other systems of modal logic, one of which we will briefly mention. Lewis reasoned that any proposition that is necessarily true must also be true. (In Leibniz's terms, if a proposition is true in *all* possible worlds, it should certainly be true in this one!) And so Lewis added as axioms to K_4 all propositions of the form $NX \supset X$. This is known as the *modal system S_4*.

The appropriate model theory for S_4 is a transitive Kripke model (but *not* a terminal one!) with the added condition that every world is accessible to itself. It is then easy to see that $NX \supset X$ holds in such a model.

The modal systems S_4 and G form a genuine parting of the ways. It is impossible to combine the two systems into a single system without getting an inconsistent system (can the reader see why?). And so we must make a choice, depending on our purpose. As a philosophical analysis of the notion of necessary truth, the system S_4 seems the appropriate one. For proof theory, the system G is the important one. But more of this in the next chapter.

Exercise 1. Why is it impossible to combine the systems G and S_4 without getting an inconsistency?

Exercise 2. In a model of type G, *no* world can be accessible to itself. Why is this?

Exercise 3. Prove that in a model of type G, there is at least one proposition p and at least one world x such that the proposition $Np \supset p$ is *false* in world x.

Exercise 4. Is the following true or false? In a Kripke model of type G, there is at least one world in which the crazy proposition N⊥ (⊥ is necessarily true) is actually true.

• 24 •

From Necessity to Provability

THE NEXT important development in modal logic after Kripke's modal semantics occurred in the early 1970s, when the *provability interpretation* was given to the word "necessary." It is surprising that this did not catch on generally sooner, since Gödel suggested it in a very short paper published in 1933. Gödel used the symbol "B" in place of Lewis's "N," and suggested that Bp be interpreted as p is *provable* (in the system of Arithmetic, or in any of the closely related systems investigated by Gödel). Now, these systems are all of type 4, and so the axioms of K_4 are all correct under that interpretation. However, the mathematical systems under investigation turned out to be even of type G (as discovered by Löb), hence it was only natural to invent a modal axiom system to take care of them. Thus the modal system G was born. It has been studied by several logicians, including Claudio Bernardi, George Boolos, D. H. J. de Jongh, Roberto Magari, Franco Montagna, Giovanni Sambin, Krister Segerberg, C. Smorynski, and Robert Solovay. Research concerning this system is still going on.

At this point, it will help to discuss modal axiom systems more rigorously. The symbolism of modal logic is that of propositional logic, with one new symbol added—we shall take this symbol to be "B." (We recall that Lewis used the symbol "N," and the more standard symbol these days is "□." But I prefer to use Gödel's symbol "B.")

By a *modal formula*—more briefly, a formula—we mean any expression formed according to the following rules:

(1) $\perp$ is a formula and each of the propositional variables p, q, r, . . . is a formula.

(2) If X and Y are formulas, so is $(X \supset Y)$.

(3) If X is a formula, so is BX.

What we called in Chapter 6 (page 43) a *formula* could now be called a *propositional formula.* A propositional formula is a special case of a modal formula; it is one in which the symbol B does not occur. But we will be concerned from now on with modal formulas, and these will be called simply *formulas.*

The logical connectives $\sim$, &, v, $\supset$, $\equiv$ are defined from $\supset$ and $\perp$ in the manner explained in Chapter 8.

Each modal system has its own axioms. In each of the modal systems that we will consider, one starts from the axioms and successively proves new formulas by use of the following two rules:

Rule 1 **(known as *modus ponens*).** Having proved X and $(X \supset Y)$, we can infer Y.

Rule 2 **(known as *necessitation*).** Having proved X, we can infer BX.

By a *formal proof* in the system is meant a nnite sequence of formulas (usually displayed vertically and read downward), called the *lines* of the proof, such that each line is an axiom of the system, or it comes from two earlier lines of the proof by Rule 1, or it comes from one earlier line of the proof by Rule 2. A formula X is called *provable* in the system if there is a formal proof whose last line is X.

The three systems that particularly interest us are the systems K, K_4, and G, whose axioms we review below.

Axioms of K: (1) All tautologies.

(2) All formulas of the form $B(X \supset Y) \supset (BX \supset BY)$.

Axioms of K_4: Those of K together with
(3) All formulas of the form BX⊃BBX.

Axioms of G: Those of K_4 together with
(4) All formulas of the form B(BX⊃X)⊃BX.

Remarks. Let us refer to the axioms of group (4) as the *special* axioms of G. We recall the Kripke–de Jongh–Sambin theorem of Chapter 18, which is that if a system of type 3 can prove all sentences of the form B(BX⊃X)⊃BX, then it can also prove all sentences of the form BX⊃BBX. Thus we could have alternatively taken as our axioms for G those of groups (1), (2), and (4); the formulas of group (3) would then have been derivable. In other words, if we add the axioms of (4) to those of K, rather than K_4, we would still get the full modal system G. The system G is often presented in this alternative manner.

Discussion. Knowledge of these modal systems provides information about the more usual systems of mathematics. The modal system K holds good for any mathematical system S of type 3 (if we interpret BX as "X is provable in S"). Similarly, the axiom system K_4 holds good for any system S of type 4, and the axiom system G holds for any system S of type G. Thus these modal axiom systems give useful information about provability in the more common types of mathematical systems (which are nonmodal). Computer scientists today are also interested in modal axiom systems for the following reason. Imagine a computer programmed to print out various sentences, some of which are assertions about what the computer can and cannot print. The interpretation of BX then becomes: "The computer can print X." Such computers are, so to speak, "self-referential," and are accordingly of interest to those working in artificial intelligence. We will consider such systems in a later chapter.

In much of this book we have been treating "belief" as a modality.

We started out using "B" for "believed" (by a reasoner of appropriate type). Modal logic enables one to give a unified treatment of reasoners who *believe* propositions, computers that can *print* propositions (or rather, sentences that express them), and mathematical systems that can *prove* propositions.

Sentential Modal Systems. By a modal *sentence* we shall mean a modal formula in which none of the propositional variables p, q, r, . . . occur—expressions such as $B\bot\supset\bot$, $B(\bot\supset B\bot)$. Thus modal sentences are all built from the five symbols B, $\bot$, $\supset$, (,). By a *sentential* modal system, we shall mean a modal system whose axioms are all sentences (from which it easily follows that only sentences are provable). For any modal system M, we shall let $\overline{M}$ be that system whose axioms consist of all *sentences* that are axioms of M, and whose rules of inference are the same as those of M (usually they will be modus ponens and necessitation). We will be particularly interested in the sentential modal systems $\overline{K}$, $\overline{K}_4$, and $\overline{G}$. It is not difficult to show that if M is either of these three systems, then any *sentence* provable in M is also provable in $\overline{M}$. (The reader might try this now as an exercise—the solution will be given later—Chapter 27—when we need to use this fact.)

We will return to the study of modal systems after treating the fascinating topic of *self-reference*—to which we turn in the following chapter.

• *Part X* •

THE HEART OF THE MATTER

• 25 •

A Gödelized Universe

LET US now turn to what can be called the heart of the matter —namely, self-reference. We have not yet given the reader any idea of *how* Gödel managed to construct a self-referential sentence—a sentence that asserted its own nonprovability in the system under consideration. He did this by inventing an extremely ingenious device known as *diagonalization.* In this chapter and the next, we will consider Gödel's diagonal argument in several forms.

A GÖDELIZED UNIVERSE

Let us contemplate a universe with *infinitely* many reasoners in it. There are also infinitely many propositions about this universe. More specifically:

(1) $\perp$ is one of these propositions (and, of course, is false).
(2) For any one of these propositions p and any reasoner R, the proposition that R *believes* p is one of these propositions.
(3) For any propositions p and q about the universe, $p \supset q$ is again a proposition about the universe and is true if and only if either p is false or q is true.

We henceforth use the word "proposition" to mean a proposition about the universe. We define the logical connectives $\sim$, &, v, $\equiv$ from $\supset$ and $\perp$ in the manner of Chapter 8.

For any reasoner R and any proposition p, we let Rp be the proposition that R believes p. For any reasoners R and S and any proposition p, RSp is the proposition that R believes that S believes p. If we throw in another reasoner K, KRSp is the proposition that K believes that R believes that S believes p—and similarly if we add more reasoners.

The reasoners had great fun reasoning about these propositions, but things were rather chaotic until a certain logician from another universe visited their universe and put things in order. The first thing he noticed was that the reasoners had no names, and so he assigned to each reasoner a number (a positive whole number, that is) known as the *Gödel number* of the reasoner. No two reasoners had the same number and every number was the number of some reasoner. Now the reasoners had names: R_1 is the reasoner whose number is 1, R_2 is the reasoner whose number is 2, and for each n, R_n is the reasoner whose number is n.

Next, the logician arranged all the propositions about the universe in a certain infinite sequence $p_1, p_2, \ldots, p_n, \ldots$ For each n, the number n was known as the Gödel number of the proposition p_n. After having done these things, the logician left and went back to his home universe.

Shortly after the logician's departure, the more clever of the inhabitants realized the following curious fact.

Fact 1. For each reasoner R, there was a reasoner R* such that for any proposition p_i, the reasoner R* believed p_i if and only if the reasoner R believed that R_i believed p_i. (Thus for any reasoner R, there was a reasoner R* such that for every number i, the proposition $R^*p_i \equiv RR_ip_i$ is true.)

This fact has some interesting consequences, as the following problems will reveal.

1 · The Diagonal Principle

Prove that for any reasoner R, there is at least one proposition p such that p≡Rp is true (in other words, p is true if and only if R believes p).

2 · Inept Reasoners

A reasoner of this universe is called *totally inept* if he believes all false propositions and doesn't believe any true ones.

Prove that no reasoner of this universe is totally inept.

Another important fact about this universe was realized soon after the logician's departure.

Fact 2. For any reasoner R and any proposition q, there is some reasoner S such that for any proposition p, S believes p if and only if R believes p⊃q. (Thus Sp≡R(p⊃q) is true.)

3 · A Tarskian Principle

A reasoner of this universe is called *perfect* if he believes all true propositions and doesn't believe any false ones. (He is the diametric opposite of a totally inept reasoner.)

For years the reasoners of this universe were in search of a perfect reasoner (of their own universe), but couldn't find one. Why were they unable to find one?

The next important things to be discovered about this universe are that certain propositions are called *established* and that there is a reasoner E who believes those and only those propositions that are established.

4

Assuming that all the established propositions are true, prove that the set of established propositions is incomplete—i.e., there must be at least one proposition p such that neither p nor $\sim p$ is established. (This means also that the reasoner E can never believe p and can never believe $\sim p$. He must remain forever undecided as to whether or not p is true.)

Next, the following two facts were revealed.

Fact I. For any reasoner R, there is a reasoner R* such that for every number i, the proposition $R^*p_i \equiv RR_ip_i$ is *established.* (This differs from Fact 1 in that we now say *established* instead of *true.*)

Fact II. For any reasoner R and any proposition q, there is a reasoner S such that for every number i, $Sp_i \equiv R(p_i \supset q)$ is established. (This differs from Fact 2 in that we say established instead of true.)

5

Prove that for any reasoner R, there is a proposition p such that the proposition $p \equiv Rp$ is established.

6

Suppose that the set of established propositions is of type 1.

(a) Show that for every reasoner R and every proposition q, there is a proposition p such that the proposition $p \equiv R(p \supset q)$ is established.

(b) Show that for every reasoner R and every proposition q, there is a proposition p such that the proposition $p \equiv (Rp \supset q)$ is established.

Finally, the following fact was realized.

Fact III. The reasoner E was of type 4 (and hence the set of established propositions was of type 4).

7

Prove that the set of established propositions is of type G.

We now see that the set of established propositions of this universe is of type G, and hence, if this set is consistent, the *fact* that it is consistent, though true, is not one of the established propositions of the universe. Equivalently, if the reasoner E is consistent, he can never know that he is consistent.

RELATION TO MATHEMATICAL SYSTEMS

The reader might wonder what all this has to do with the theory of mathematical systems. Well, suppose we have a mathematical system S with all its propositions arranged in some infinite sequence p_1, p_2, . . . , p_n, . . . Now, instead of *reasoners,* we will consider certain *properties* of propositions; these properties are also arranged in some infinite sequence R_1, R_2, . . . , R_n, . . . For any *property* R_i and any proposition p_j, we now let R_ip_j be the proposition that the property R_i *holds* for the proposition p_j. Suppose also that the property—call it E—of being a *provable* proposition of the system is one of the properties of the above list and suppose that Facts I, II, and III hold replacing the words "reasoner" by "property" and "established" by "provable" (provable in S, that is). Then by a mere change of terminology, the preceding arguments of this chapter show that the system S must be of type G.

Actually, the systems investigated by Gödel did not start out with properties of propositions, but with properties of *numbers.* However, by the device of Gödel-numbering the propositions, any property of numbers then corresponds to a certain property of propositions—

namely, for any property A of *numbers,* we let A′ be the property which holds for just those *propositions* p_i such that A holds for the number i. A concrete illustration of this will be given in the next chapter.

Self-reference can also be achieved without Gödel numbering, as is indicated by the exercise below.

Exercise 1. There is another universe of reasoners in which it makes no difference whether the number of reasoners is finite or infinite. Some of the reasoners are immortal, but no reasoner knows which of the reasoners are immortal and which ones are mortal. In fact, no reasoner knows whether he himself is mortal or not. For every reasoner R, we let $\overline{R}$ be the proposition that R is immortal. For any reasoners R and S, we let $R\overline{S}$ be the proposition that R *believes* that S is immortal; for any three reasoners R, S, and K, we let $RS\overline{K}$ be the proposition that R believes that S believes that K is immortal —and so forth (if there are more reasoners involved).

In place of Fact 1 of the last universe, we have the following fact for this universe: For any reasoner R, there is a reasoner R* such that for every reasoner S, the reasoner R* believes that S is immortal if and only if R believes that S believes himself to be immortal (thus $R^*\overline{S}$ is true if and only if $RS\overline{S}$ is true).

Given a reasoner R, find a proposition p such that p is true if and only if R believes p.

Note: This method of achieving self-reference without Gödel numbering is borrowed from the field known as combinatory logic. A host of related self-referential problems in this field can be found in my book *To Mock a Mockingbird.*

SOLUTIONS

1 · Take any reasoner R. By Fact 1, there is a reasoner R* such that for every number i, the reasoner R* believes p_i if and only if R

believes the proposition $R_i p_i$. Now, R* has some Gödel number h, and so R* is the reasoner R_h. Thus for any number i, the following is true:

(1) R_h believes p_i if and only if R believes that R_i believes p_i.

Since this is true for *every* number i, it is also true when i is the number h. We thus have:

(2) R_h believes p_h if and only if R believes that R_h believes p_h.

We thus take p to be the proposition that R_h believes p_h, and we see that p is true if and only if R believes p.

2 · We have just seen that for any reasoner R, there is a proposition p such that p is true if and only if R believes p. This means that one of the following two cases must hold: (1) p is true and R believes p; (2) p is false and R doesn't believe p. If (1) holds, then R believes at least one true proposition—namely, p—and hence R is not totally inept. If (2) holds, then there is at least one false proposition—namely, p—such that R doesn't believe p, hence R doesn't believe *all* false propositions, and so again R is not totally inept.

3 · Using Fact 2, we take for q the proposition $\perp$. Then for any reasoner R, there is a reasoner R′ (called "S," in Fact 2) who believes those and only those propositions p such that R believes $p \supset \perp$. (Such a reasoner R′ might be said to *oppose* R.)

Now, suppose R were perfect. Then for any proposition p, R believes $p \supset \perp$ if and only if $p \supset \perp$ is true, which in turn is the case if and only if p is false. Therefore R believes $p \supset \perp$ if and only if p is false. Also, R′ believes p if and only if R believes $p \supset \perp$. Putting these last two facts together, R′ believes p if and only if p is false. This means that R′ is totally inept.

We thus see that if the universe contains a perfect reasoner R, then it must also contain a totally inept reasoner R′. But we have proved in Problem 2 that the universe contains no totally inept reasoner. Therefore the universe contains no perfect reasoner.

4 · Let us call a reasoner *accurate* if he believes no false propositions. Consider now any accurate reasoner R. We saw in Problem 3 that none of the reasoners is perfect, hence R is also imperfect. This means that R either believes some false proposition, or he fails to believe some true proposition. Since R is accurate, he doesn't believe any false proposition, hence it must be that R fails to believe some true proposition. This proves that for any accurate reasoner R, there is at least one true proposition p that R fails to believe. Since p is true, $\sim$p is false, hence R, being accurate, doesn't believe $\sim$p either. Therefore, for every accurate reasoner R, his belief system is incomplete. There is at least one proposition p such that R neither believes p nor believes $\sim$p—he must remain forever undecided as to whether p is true or false.

Assuming that all the established propositions are true, the reasoner E is accurate (because he believes all the established propositions and no others). Therefore there is a proposition p such that E neither believes p nor believes $\sim$p, hence neither p nor $\sim$p is established.

5 · The proof is essentially that of Problem 1 using Fact I in place of Fact 1.

Given a reasoner R, there is a reasoner R_h (called R*) such that for any number i, the proposition $R_hp_i \equiv RR_ip_i$ is established. Hence $R_hp_h \equiv RR_hp_h$ is established. Thus $p \equiv Rp$ is established, where p is the proposition R_hp_h.

6 · Suppose the set of established propositions is of type 1.

(a) Take any reasoner R and any proposition q. By Fact 2, there is a reasoner S such that for every p, the proposition $Sp \equiv R(p \supset q)$ is established. Also by Problem 5 (reading S for R), there is a proposition p such that $p \equiv Sp$ is established. It then follows that $p \equiv R(p \supset q)$ is established (since it is a logical consequence of the last two propositions).

(b) Again take any reasoner R and any proposition q. By (a), there

is a proposition—call it p_1—such that $p_1 \equiv R(p_1 \supset q)$ is established. Hence $(p_1 \supset q) \equiv (R(p_1 \supset q) \supset q)$ is established (a trick we have used before), and so $p \equiv (Rp \supset q)$ is established, where p is the proposition $p_1 \supset q$.

7 · We are now given that E is of type 4, hence he is certainly of type 1. Then by (b) of the last problem, for any proposition q, there is a proposition p such that the proposition $p \equiv (Ep \supset q)$ is established —thus the set of established propositions is *reflexive.* And we know from Chapter 19 that any reflexive system of type 4 is of type G.

• 26 •

Some Remarkable Logic Machines

FERGUSSON'S LATEST MACHINE

The logician Malcolm Fergusson of my book *The Lady or the Tiger?* was fond of constructing logic machines to illustrate important principles in logic and proof theory. One of his machines was described in that book. In Fergusson's later years, when he heard about Gödel's and Löb's theorems, he straightaway set out to construct a second machine, which he delighted in demonstrating to his friends. He proved to their satisfaction that the machine was a consistent and stable machine of type G, and he took particular delight in demonstrating that the machine, though consistent, could never prove its own consistency! The machine illustrates in a very simple and instructive manner the essential ideas behind Gödel's First and Second Incompleteness theorems as well as Löb's Theorem. I am accordingly happy to communicate the details to the reader.

The machine prints out various sentences built from seventeen symbols. The first seven symbols are the following:

P	⊥	⊃	(	)	d	,
1	2	3	4	5	6	7

Underneath each of these seven symbols, I have written its *Gödel number.* The remaining ten symbols are the familiar digits 1, 2, 3, 4, 5, 6, 7, 8, 9, 0. These digits are assigned Gödel numbers as follows: The Gödel number of 1 is 89 (8 followed by one 9); the Gödel number of 2 is 899 (8 followed by two 9's); and so on, until 0, whose Gödel number is 89999999999 (8 followed by ten 9's). Thus each of the seventeen symbols has a Gödel number. Given a complex expression, one finds its Gödel number by replacing each symbol with its Gödel number—for example, the Gödel number of (P⊥⊃⊥) is 412325. As another example, the Gödel number of P35 is 18999899999. For any expression E, by $\overline{E}$ we shall mean the Gödel number of E (written as a string of the digits 1, 2, . . . , 0). Not every number is the Gödel number of an expression (for example, 88 is not the Gödel number of any expression). If n *is* the Gödel number of an expression, we shall sometimes refer to the expression as the n^{th} expression. (For example, pd is the 16^{th} expression; ⊥ is the 2^{nd} expression.)

The machine is *self-referential* in that the sentences printed by the machine express propositions about what the machine can and cannot print. An expression is called *printable* if the machine can print it. The symbol "P" means "printable," and for any expression E built from the seventeen symbols, if we want to write down a sentence that asserts that E is printable, we don't write down PE, but $P\overline{E}$ (i.e., P followed by the Gödel number of E). For example, a sentence that asserts that (P⊥⊃⊥) is printable is $P\overline{(P\bot\supset\bot)}$—i.e., P412325.

For any expressions X and Y, Fergusson defined the *diagonalization of X with respect to Y* to be the expression $(X(\overline{X},\overline{Y})\supset Y)$. The symbol "d" abbreviates "diagonalization"—and for any expressions X and Y, the expression $Pd(\overline{X},\overline{Y})$ is a sentence expressing the proposition that the diagonalization of X with respect to Y is printable.

We shall now define what it means for an expression to be a *sentence* and what it means for a sentence to be *true.*

(1) $\perp$ is a sentence and $\perp$ is false.
(2) For any expression X, the expression $P\overline{X}$ is a sentence and is true if and only if the expression X is printable.
(3) For any expressions X and Y, the expression $Pd(\overline{X},\overline{Y})$ is a sentence and is true if and only if the expression $(X(\overline{X},\overline{Y})\supset Y)$—which is the diagonalization of X with respect to Y—is printable.
(4) For any sentences X and Y, the expression $(X\supset Y)$ is a sentence and is true if and only if either X is not true or Y is true.

It is to be understood that no expression is a sentence unless its being so is a consequence of the above rules. The logical connectives $\sim$, &, v, $\equiv$ are defined from $\supset$ and $\perp$ in the manner explained in Chapter 8.

Now we give the rules for what the machine can print. The machine is programmed to print out an infinite list of sentences sequentially. Certain sentences called *axioms* can be printed at any stage of the process. Among the axioms are all tautologies. (Thus for any tautology X, the machine can print X whenever it likes, regardless of what it has or has not printed at any previous stage.) Next, the machine is programmed so that for any sentences X and Y, if the machine has, at a certain stage, already printed X and $X\supset Y$, then it can print Y. Thus the machine is of type 1 (in the sense that the class of printable sentences is of type 1). Since it is true that if X and $X\supset Y$ are both printable, so is Y, then the sentence $(P\overline{X}\&P(\overline{X\supset Y}))\supset P\overline{Y}$ is true; or, what is the same thing, the sentence $P(\overline{X\supset Y})\supset(P\overline{X}\supset P\overline{Y})$ is true (the two sentences are logically equivalent). Well, the machine "knows" the truth of all sentences of the form $P(\overline{X\supset Y})\supset(P\overline{X}\supset P\overline{Y})$ and takes them all as axioms. Thus the machine is of type 2. Next, if the machine ever prints a sentence X, it "knows" that it has printed X and will sooner or later print the true sentence $P\overline{X}$. (The sentence $P\overline{X}$ *is* true, since X has been printed.) And so the machine is normal, hence is of type 3. Since the machine is normal, then for any sentence X, the sentence $P\overline{X}\supset P\overline{P\overline{X}}$ is true. So the machine is initially "aware" of the truth of

all such sentences and takes them all as axioms. Thus the machine is of type 4.

There is one more thing the machine can do, and this is quite crucial. For any expressions X and Y, the sentence $Pd(\overline{X},\overline{Y})$ is true if and only if $(X(\overline{X},\overline{Y})\supset Y)$ is printable, which in turn is the case if and only if the sentence $P\overline{(X(\overline{X},\overline{Y})\supset Y)}$ is true. Therefore the following sentence is true: $Pd(\overline{X},\overline{Y})\equiv P\overline{(X(\overline{X},\overline{Y})\supset Y)}$.

Well, the machine knows the truth of all such sentences and takes them all as axioms. These axioms are called the *diagonal axioms.*

Let us now systematically review all the axioms and operations of the machine:

Axioms: *Group 1.* All tautologies.

Group 2. All sentences of the form $P\overline{(X\supset Y)}\supset(P\overline{X}\supset P\overline{Y})$.

Group 3. All sentences of the form $P\overline{X}\supset P\overline{P\overline{X}}$.

Group 4 (the diagonal axioms). All sentences of the form $Pd(\overline{X},\overline{Y})\equiv P\overline{(X(\overline{X},\overline{Y})\supset Y)}$, where X and Y are any expressions (not necessarily sentences).

Operation Rules. (1) Axioms can be printed at any stage.

(2) Given sentences X and $(X\supset Y)$ already printed, the machine can then print Y.

(3) Given a sentence X already printed, the machine can print $P\overline{X}$.

This concludes the rules governing printability by the machine. It is to be understood that the only way the machine can print a sentence X at a given stage is by following one of the above rules. That is, X is printable at a given stage *only* if one of the following three conditions holds: (1) X is an axiom; (2) there is a sentence Y such that Y and $(Y\supset X)$ have both been printed at an earlier stage; (3) there is a sentence Y such that X is the sentence $P\overline{Y}$ and Y has been printed at an earlier stage.

Remarks. For each sentence X, let BX be the sentence $P\overline{X}$. The symbol "B" is *not* part of the machine language; we are using it to

talk about the machine. We are using "B" as standing for the operation that assigns to each sentence X the sentence $P\overline{X}$. When we say that the machine is of type 4, we mean that it is of type 4 with reference to this Operation B. In essence, without the diagonal axioms, the axiom system of this machine is the modal system K_4. We will shortly see that adding the diagonal axioms gives us all the power of the modal system G.

Provability. We have defined for each sentence what it means for the sentence to be *true,* and so each sentence expresses a definite proposition, which might be true or might be false. We say that the machine *proves* a proposition if it prints some sentence that expresses the proposition. For example, the sentence $\sim P2$ expresses the proposition that the machine is consistent (since 2 is the Gödel number of $\bot$), and so if the machine printed $\sim P2$, it would prove its own consistency. If the machine printed P2, then it would prove its own *inconsistency.*

We say that the machine is *accurate* if all propositions provable by the machine are true. We say that the machine is *consistent* if it cannot prove $\bot$, and that the machine is *stable* if for every sentence X, if $P\overline{X}$ is printable, so is X.

Reflexivity. Now we turn to the proof that the machine is Gödelian —in fact, reflexive.

1 · The Gödel Sentence G

Find a sentence G such that the sentence $G \equiv \sim P\overline{G}$—i.e., the sentence $G \equiv (P\overline{G} \supset \bot)$—is printable.

2 · Reflexivity

Show that for any sentence Y, there is a sentence X such that the sentence $X \equiv (P\overline{X} \supset Y)$ is printable.

Solutions. Problem 1 is a special case of Problem 2, so we first solve Problem 2. Let Y be any sentence. For any expression Z, the sentence $Pd(\overline{Z},\overline{Y})\equiv P\overline{(Z(\overline{Z},\overline{Y})\supset Y)}$ is printable (because it is one of the diagonal axioms). We take for Z the expression Pd, and therefore $Pd(\overline{Pd},\overline{Y})\equiv P\overline{(Pd(\overline{Pd},\overline{Y})\supset Y)}$ is printable. Since the machine is of type 1, it then follows that the following sentence is printable: $(Pd(\overline{Pd},\overline{Y})\supset Y)\equiv(P\overline{(Pd(\overline{Pd},\overline{Y})\supset Y)}\supset Y)$.

Thus the sentence $X\equiv(P\overline{X}\supset Y)$ is printable, where X is the sentence $(Pd(\overline{Pd},\overline{Y})\supset Y)$.

Problem 1 is a special case of 2, taking $\perp$ for Y. Thus the Gödel sentence G for this machine is $Pd(\overline{Pd},\overline{\perp})\supset\perp$—i.e., the sentence $(Pd(16,2)\supset\perp)$.

Let us take a closer look at Gödel's remarkable sentence G. First, what does the sentence Pd(16,2) say? It says that the diagonalization of the 16th expression with respect to the 2nd expression is printable. Now, the 16th expression is Pd and the 2nd expression is $\perp$, and so Pd(16,2) says that the diagonalization of Pd with respect to $\perp$ is printable, but this diagonalization is the sentence $(Pd(16,2)\supset\perp)$—i.e., the very sentence G! And so Pd(16,2) says that G *is* printable, hence $(Pd(16,2)\supset\perp)$—which is the sentence G—says that G is *not* printable (or, what is the same thing, that the printability of G implies falsehood). Thus G says that G is not printable; G is *true* if and only if G is not printable. Thus G asserts its own nonprintability. Here, in a nutshell, is Gödel's ingenious method of achieving self-reference.

The sentence $G\equiv\sim P\overline{G}$—i.e., the sentence $G\equiv(P\overline{G}\supset\perp)$—is not only true, but actually printable (it is one of the diagonal axioms). Since the machine is normal and of type 1, it follows by Gödel's First Incompleteness Theorem (Theorem I, Chapter 20, page 174) that if the machine is consistent, then G is not printable, and if the machine is also stable, then $\sim$G is also not printable. And so, *if* the machine is both consistent and stable, the sentence G is undecidable in the system of sentences that the machine can print.

Now, the machine is in fact of type 4, and since it is Gödelian

—the sentence $G \equiv \sim P\overline{G}$ is printable—it then follows from Gödel's Second Incompleteness Theorem (Part 4 of Summary I*, Chapter 13, page 110) that if the machine is consistent, then it cannot prove its own consistency—i.e., it cannot print the sentence $\sim P2$. Also, if the machine is consistent, then the sentence $\sim P2$ is *true,* and is hence another example of a true sentence that the machine cannot print.

Furthermore, the machine is reflexive (Problem 2) and being of type 4, it must be Löbian (by Löb's Theorem), and so for any sentence X, if $P\overline{X} \supset X$ is printable, so is X. It then follows by Theorem M_1, Chapter 18 (page 157), that the machine is of type G.

THE CORRECTNESS OF THE MACHINE

We have shown that *if* Fergusson's machine is consistent, then it cannot prove its own consistency; but how do we know whether or not the machine is consistent? We will now prove that the machine is not only consistent, but wholly accurate—i.e., that every sentence printed by the machine is true.

We have already shown that all the *axioms* of the machine are true, but let us carefully review the arguments: The axioms of Group 1 are all tautologies, hence they are certainly true. As for the axioms of Group 2, to say that $P\overline{(X \supset Y)} \supset (P\overline{X} \supset P\overline{Y})$ is true is to say that if $P\overline{(X \supset Y)}$ and $P\overline{X}$ are both true, so is PY—which is to say that if $(X \supset Y)$ and X are both printable, so is Y. Well, this is obviously the case by virtue of Operation 2. Thus the axioms of Group 2 are all true. As for the axioms of Group 3, to say that $P\overline{X} \supset PP\overline{X}$ is true is to say that if $P\overline{X}$ is true, so is $PP\overline{X}$, which in turn is to say that if X is printable, so is $P\overline{X}$—and this is indeed the case, by virtue of Operation 3. As for the diagonal axioms, $Pd(\overline{X},\overline{Y})$ is true if and only if $(X(\overline{X},\overline{Y}) \supset Y)$ is printable, which is the case if and only if $P\overline{(X(\overline{X},\overline{Y}) \supset Y)}$ is true. Therefore $Pd(\overline{X},\overline{Y}) \equiv P\overline{(X(\overline{X},\overline{Y}) \supset Y)}$ is true.

Now we know that all the axioms of the machine are true, but we need to show that all the *printable* sentences are true.

We recall that the machine prints sentences at various *stages.* We now wish to establish the following lemma, theorem, and corollary.

Lemma. If X is a sentence printed at a certain stage and all sentences printed prior to that stage are true, then X is also true.

Theorem 1. Every sentence printed by the machine is true.

Corollary. The machine is both consistent and stable.

3

How are the above lemma, theorem, and corollary proved?

Solutions. First we prove the lemma: Assume the hypothesis that all sentences previously printed are true; we are to show that X is true.

Case 1. X is an axiom. Then X is true (as we have already proved).

Case 2. There is a sentence Y such that Y and $(Y \supset X)$ have already been printed. Then by the assumed hypothesis, Y and $(Y \supset X)$ are both true, hence X must be true.

Case 3. X is of the form $P\overline{Y}$, where Y is a sentence that has been previously printed. Since Y has been printed, then $P\overline{Y}$ is true—i.e., X is true.

This concludes the proof of the lemma.

Proof of Theorem 1. The machine is programmed to print all printable sentences in some definite infinite sequence X_1, X_2, . . . , X_n, . . . By X_n is meant the sentence printed at stage n. Now, the first sentence printed by the machine (the sentence X_1) must be an axiom (since the machine hasn't printed any other sentences yet), hence X_1 must be true. If the above infinite list should contain

any false sentence, then there must be a *smallest* number n such that X_n is false—that is, there must be a *first* false sentence that the machine prints. We know that n is not equal to 1 (since X_1 is true), hence n is greater than 1. This means that the machine prints a false sentence at stage n but has printed only true sentences at all earlier stages. But this is contrary to the lemma. Therefore the machine can never print any false sentences.

Proof of Corollary 1. Since the machine is accurate (by Theorem 1), then ⊥ can never be printed, because ⊥ is false. Therefore the machine is consistent.

Next, suppose that $P\overline{X}$ is printable. Then $P\overline{X}$ is true (by Theorem 1), which means that X is printable. Therefore the machine is stable.

We now see that Fergusson's machine *is* consistent, but can never prove its own consistency. Thus you and I (as well as Fergusson) know that the machine is consistent, but the poor machine doesn't have that knowledge!

CRAIG'S VARIANT

When Inspector Craig (a good friend of Fergusson's) heard of Fergusson's machine, he thought of an interesting variant, which does not involve Gödel numbering. Craig's machine used just the following six symbols:

P ⊥ ⊃ ( ) R

His definitions of *sentence* and *true sentence* are given by the following rules:

(1) ⊥ is a sentence and ⊥ is not true.
(2) For any sentences X and Y, (X⊃Y) is a sentence and is true if and only if X is not true or Y is true.

(3) For any sentence X, PX is true if and only if the machine can print X.

(4) For any sentences X and Y, the expression (XRY) is a sentence and is declared true if and only if the machine can print ((XRX)⊃Y).

At first sight it might appear that (4) is circular, since the truth of (XRY) is defined in terms of an expression involving the letter R; but this circularity is only apparent. If Craig had defined (XRY) to be true if and only if ((XRX)⊃Y) is *true,* the definition would have been circular, since we couldn't know what it means for (XRY) to be true without first knowing what it means for (XRX) to be true. But Craig didn't do that. The fact is that for any sentences X and Y, either the expression ((XRX)⊃Y) is printable or it isn't. The symbol "R" stands for the relation that holds between X and Y when ((XRX)⊃Y) *is* printable. Thus Craig is not defining the relation R in terms of itself, but in terms of the *symbol* "R," and this constitutes no circularity.

The first three groups of axioms of Craig's machine are the same as those of the modal system K_4 (where *sentence* means sentence in the machine language of Craig's system). Thus they are like the first three groups of axioms of Fergusson's machine, leaving out the bars over X, Y, Z. The fourth group of Craig's axioms can be read as: (4)′ (Craig's diagonal axioms). All sentences of the form (XRY)≡P((XRX)⊃Y).

Of course all of Craig's diagonal axioms are true sentences.

The operation rules of Craig's machine are the same as those of Fergusson's machine.

4

(a) Prove that Craig's machine is reflexive (and hence also of type G, since it is of type 4).

(b) Find a Gödel sentence for Craig's machine (i.e., a sentence G such that G≡~PG is printable by Craig's machine).

Solution. (a) Since for any sentences X and Y, the sentence (XRY)≡P((XRX)⊃Y) is printable (it is a diagonal axiom), then this is also the case if X happens to be the very sentence Y. Therefore (YRY)≡P((YRY)⊃Y) is printable, hence so is the sentence ((YRY⊃Y)≡(P(YRY)⊃Y)⊃Y), and thus Z≡(PZ⊃Y) is printable, where Z is the sentence ((YRY)⊃Y).

(b) Taking ⊥ for Y, we get the Gödel sentence ((⊥R⊥)⊃⊥).

Note: The axioms of Craig's machine are also all true, and by the same reasoning we used for Fergusson's machine, it can be seen that every sentence printable by Craig's machine is true. Therefore Craig's machine is also consistent and stable (and of type G).

5 · McCulloch's Observation

When Walter McCulloch (a friend of both Craig and Fergusson) was informed about Craig's machine, he made the following interesting observation: Given any sentences X and Y in which the symbol "R" does not occur, there is a sentence Z in which "R" does not occur such that the sentence (XRY≡Z) is printable. (This implies, incidentally, that for any sentence X at all, there is a sentence X′ in which "R" does not occur such that X≡X′ is printable by Craig's machine.)

Can you prove that McCulloch's observation is correct?

Solution. We proved in the solution of Problem 8 of Chapter 19 that any system of type G which can prove p≡B(p⊃q) can prove p≡Bq. (We showed this for reasoners, but the same argument goes through for systems.) Now, Craig's system is of type G, and for any sentence X, the sentence (XRX)≡P((XRX)⊃X) is printable, and so

if we take (XRX) for p and X for q, it follows that (XRX)≡PX is printable. From this it follows that ((XRX)⊃Y)≡(PX⊃Y) is printable, and hence (by regularity) P((XRX)⊃Y)≡P(PX⊃Y) is printable. But also (XRY)≡P((XRX)⊃Y) is printable, hence (XRY)≡P(PX⊃Y) is printable. If now the symbol "R" doesn't occur in either X or Y, then it doesn't occur in P(PX⊃Y), and so we take Z to be P(PX⊃Y).

• 27 •

Modal Systems Self-Applied

INSPIRED BY the logic machines of Craig and Fergusson, I would like now to look at modal axiom systems from the viewpoint of *self-referential interpretations.*

We recall that by a *modal sentence*—more briefly, a *sentence*—we mean a modal formula without propositional variables. We now define a modal sentence to be *true* for a modal system M if it is true when we interpret B as *provable in M.* Thus:

(1) $\perp$ is false for M.

(2) For any modal sentences X and Y, the sentence $X \supset Y$ is true for M if and only if either X is *not* true for M or Y *is* true for M.

(3) For any sentence X, the sentence BX is true for M if and only if X is *provable* in M.

Note: A sentence X can be *true* for a modal system M without being provable in M, and conversely. For example, the sentence $\sim B\perp$ is *true* for M if and only if M is consistent. To say that $\sim B\perp$ is *provable* in M is to say that M can prove its own consistency. We will soon see that the modal system G is consistent, and so the sentence $\sim B\perp$ is *true* for G. But the sentence $\sim B\perp$ is not *provable* in G. On the other hand, any inconsistent modal system of type 1 can prove all sentences, hence in particular the sentence $\sim B\perp$. In this case the sentence $\sim B\perp$ is *provable* in the system, but not *true* for the system.

Consider now a modal system M_1 and a modal system M_2 (possibly the same as M_1 or possibly different). We shall say that M_1 is *correct* for M_2 if every sentence provable in M_1 is *true* for M_2. And we shall say that a modal system M is *self-referentially correct* if M is correct for M—in other words, if every sentence provable in M is true for M.

Any self-referentially correct system must be consistent (because if $\perp$ were provable in the system, the system couldn't be self-referentially correct, since $\perp$ is false for the system) and must also be stable (because if BX is provable in the system and the system is self-referentially correct, then BX must be true for the system, which means that X is provable in the system). And so any self-referentially correct system is automatically both consistent and stable.

Self-referential correctness has one curious feature. It is possible that a system M might be self-referentially correct; yet if some of the axioms were deleted, the resulting system might no longer be self-referentially correct. For example, we might have one axiom that asserts that a second axiom is provable in the system; if this second axiom is removed, the first axiom might become false!

Some Self-Referentially Correct Systems. We recall the sentential modal systems $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ described in Chapter 24 (they are like the systems K, K_4, and G, except that the axioms are all restricted to sentences). We now aim to show that these three systems are self-referentially correct (from which, incidentally, we will be able to show that the systems K, K_4, and G are self-referentially correct).

If $\overline{M}$ is any of these three axiom systems $\overline{K}$, $\overline{K_4}$, $\overline{G}$, to show that $\overline{M}$ is self-referentially correct, it suffices to show that all *axioms* of $\overline{M}$ are true for $\overline{M}$—the reason for this is a consequence of the following lemma, which has other applications as well.

Lemma A. Let $\overline{M_1}$ be any sentential modal system whose only inference rules are modus ponens (from X and X$\supset$Y we can infer Y)

and the necessitation rule (from X we can infer BX). Let M_2 be any modal system such that all sentences provable in $\overline{M_1}$ are provable in M_2 and such that all *axioms* of $\overline{M_1}$ are *true* for M_2. Then all sentences provable in $\overline{M_1}$ are true for M_2—i.e., $\overline{M_1}$ is *correct* for M_2.

Corollary A1. For any sentential modal system $\overline{M}$ whose only inference rules are modus ponens and necessitation, if all *axioms* of $\overline{M}$ are true for $\overline{M}$, then $\overline{M}$ is self-referentially correct.

1

Prove Lemma A. (Hint: The proof is essentially the same as the argument we gave in the last chapter to show that all provable sentences of Fergusson's machine were true—once we established that all the axioms of the machine were true.)

Solution. Consider any sequence $X_1, \ldots, X_n$ of sentences that constitutes a proof in the system $\overline{M_1}$. We will see that the first line X_1 must be true for M_2, then the second line X_2 must be true for M_2, then the third line X_3, and so forth down to the last line.

The first line X_1 must be an axiom, hence it is true for M_2 by hypothesis. Now consider the second line X_2. Either it is an axiom of $\overline{M_1}$ (in which case it is true for M_2), or it must be the sentence BX_1, in which case it is certainly true for M_2 since X_1 has already been proved in M_1, and is hence provable in M_2. Now we know that the first two lines are true for M_2. We now consider the third line X_3. If it is either an axiom of $\overline{M_1}$ or is of the form BY, where Y is an earlier line (X_1 or X_2), then X_3 is true for M_2 for the same reasons as before. If X_3 is neither, then it must be derived from X_1 and X_2 by modus ponens, and since we already know that X_1 and X_2 are true for M_2, it follows that X_3 is true for M_2. (Clearly, for any sentences X and Y, if X and $X \supset Y$ are both true for M_2, then Y is true for M_2.) Now we know that the first three lines of the proof are all true for M_2, and knowing this, the truth of X_4 can be

established by the same argument. Then, knowing the truth of the first four lines, we similarly get the truth of the fifth line—and so on, until we reach the last line.† This concludes the proof of Lemma A.

The self-referential correctness of the systems $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ follows from Corollary A_1 and the following lemma.

Lemma B. For any modal system M:

(a) If M is of type 1, then all axioms of $\overline{K}$ are true for M.

(b) If M is normal and of type 1, then all axioms of $\overline{K_4}$ are true for M.

(c) If M is of type G, then all axioms of $\overline{G}$ are true for M.

2

Why is Lemma B correct?

Solution. (a) Suppose that the set of provable sentences of M is closed under modus ponens. All tautologies are obviously true for M (they are true for any modal system whatsoever). The other axioms of $\overline{K}$ are sentences of the form $(BX\&B(X\supset Y))\supset BY$—or alternatively $B(X\supset Y)\supset(BX\supset BY)$; it really makes no difference. To say that $(BX\&B(X\supset Y))\supset BY$ is true for M is to say that if X is provable in M and $X\supset Y$ is provable in M, so is Y. Well, this is the case, since the provable sentences of M are closed under modus ponens.

(b) Suppose M is a normal system of type 1. Then all axioms of $\overline{K}$ are true for M—by (a). The other axioms of $\overline{K_4}$ are the sentences of the form $BX\supset BBX$. Well, to say that such a sentence is true for M is to say that if BX is true for M, so is BBX—in other words, if X is provable in M, so is BX. This is so, since M is normal.

(c) Suppose M is of type G. Then it is certainly of type 4, so by (b), all axioms of $\overline{K_4}$ are true for M. The remaining axioms of $\overline{G}$ are

†This type of argument is known as *mathematical induction.*

sentences of the form $B(BX \supset X) \supset BX$, and to say that it is *true* for M is to say that if $BX \supset X$ is provable in M, so is X—in other words, that M is Löbian. Well, we proved in Chapter 19 that any system of type G is Löbian. This concludes the proof of Lemma B.

We continue to let M be any of the systems K, K_4, or G. Then $\overline{M}$ is respectively $\overline{K}$, $\overline{K_4}$, or $\overline{G}$.

Corollary B_1. All axioms of $\overline{M}$ are true for $\overline{M}$.

Corollary B_2. All axioms of $\overline{M}$ are true for M.

Proofs. Since $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ are respectively of type 1, normal and of type 1, and of type G, Corollary B_1 is immediate from Lemma B. Also the systems K, K_4, and G are respectively of type 1, normal and of type 1, and of type G, and so we also have Corollary B_2.

From Corollary B_1 and Corollary A_1, we now have Theorem 1.

Theorem 1. The systems $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ are all self-referentially correct.

Corollary. The systems $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ are consistent and stable.

We now have a third example of a consistent and stable system of type G—namely, the modal system $\overline{G}$ (the other two systems being the machines of the last chapter). We will soon see that the modal system G is also self-referentially correct (hence also both consistent and stable).

I hope the reader now fully realizes the absurdity of doubting the consistency of a system on the mere grounds that it cannot prove its own consistency!

The Systems K, K_4, and G. To establish the self-referential correctness of the systems K, K_4, and G, we proceed as follows. First of all, Lemma A has another corollary.

Corollary A_2. Let M be any modal system whose only inference rules are modus ponens and necessitation. Then, if all axioms of $\overline{M}$ are true for M, all provable sentences of $\overline{M}$ are true for M.

It follows from Corollary A_2 and Corollary B_2 that if M is any of the modal systems K, K_4, or G, then all sentences provable in $\overline{M}$ are true for M. But we are not quite done. It remains to show that if M is either of these three systems, then any sentence provable in M is also provable in $\overline{M}$ (a fact that was asserted without proof at the end of Chapter 19). Once this is done, the proof of the self-referential correctness of K, K_4, and G will be complete.

3

Why is it true that if a sentence is provable in M (M being either K, K_4, or G), then it is provable in $\overline{M}$?

Solution. One can see by inspection of either of these three systems that if X is an axiom of M, then if we substitute any sentences for the propositional variables in X (substituting the same sentence for different occurrences of the same variable, of course), then the resulting sentence is also an axiom of M—and hence of $\overline{M}$. We now take any one particular sentence, say T, and for any formula X, let X′ be the result of substituting T for *all* the propositional variables in X. Of course if X is itself a sentence, we take X′ to be X. We now note the following facts: (1) If X is an axiom of M, then X′ is an axiom of $\overline{M}$. (2) For any formulas X and Y, the sentence $(X \supset Y)'$ is the sentence $X' \supset Y'$, and therefore for any formulas X, Y, and Z, if Z is derivable from X and Y by modus ponens, then Z′ is derivable from X′ and Y′ by modus ponens. (3) For any formula X, the sentence $(BX)'$ is the sentence BX′ (i.e., B followed by X′), and so if Y is derivable from X by the necessitation rule, then Y′ is derivable from X′ by the necessitation rule. It therefore follows that given any sequence $X_1, \ldots, X_n$ of formulas, if this sequence

constitutes a proof in the system M, then the sequence X_1', X_2', . . . , X_n' constitutes a proof in $\overline{M}$. And so, if X is any formula provable in M, the sentence X′ is then provable in $\overline{M}$. If, furthermore, X happens to be a sentence, then $X' = X$, and hence X itself is provable in $\overline{M}$. This proves that any sentence provable in M is provable in $\overline{M}$.

• *Part XI* •

FINALE

• 28 •

Modal Systems, Machines, and Reasoners

IN THE next chapter we will meet up with some very strange reasoners indeed. To appreciate them fully, let us first turn to the topic of minimal reasoners.

MINIMAL REASONERS OF VARIOUS TYPES

A modal sentence X is in itself neither true nor false; it only expresses a definite proposition once the symbol "B" is given an interpretation. We have defined a sentence to be *true* for a modal system M if it is true when "B" is interpreted as provability in M. We have all along understood a modal sentence as being true for a *reasoner* if it is true when "B" is interpreted as *believed by the reasoner.*

To say that a sentence is *true* for a reasoner means something entirely different than saying that it is *believed* by the reasoner. For example, to say that $\sim B\bot$ is true for a reasoner is to say that the reasoner is consistent, whereas to say that $\sim B\bot$ is believed by a reasoner is to say that the reasoner believes that he is consistent.

A machine can easily be programmed to print out all and only those sentences provable in the modal system $\overline{G}$ by giving the machine these instructions: (1) At any stage, you may print any axiom of $\overline{G}$. (2) If at any stage you have printed sentences X and $(X \supset Y)$, you may then print Y. (3) If at any stage you have printed X, you may then print BX. (The machine can then be given further instructions that will guarantee that anything the machine *can* do, it sooner or later *will* do, and so every sentence provable in $\overline{G}$ will eventually be printed by the machine.) Let us call such a machine a $\overline{G}$ machine.

Now, let us imagine a reasoner with his eye constantly on the output of the machine. However, he does not interpret BX as "X is printable by the machine," or as "X is provable in $\overline{G}$," but as "I believe X." (He thinks that the machine is printing sentences about *him!*) He gives what we might call an *egocentric* interpretation of modal sentences.

Then, suppose that the reasoner has complete confidence that the machine knows what he believes, and so whenever the machine prints a sentence X, he immediately believes it (under the egocentric interpretation, of course). His belief system will then include *all* sentences provable in $\overline{G}$. This does *not* guarantee that he is of type G (he may not be normal, even though he believes he is, and he may not even be of type 1, even though he believes he is). Of course if he *correctly* believes all sentences provable in $\overline{G}$, then it is easy to see that he is of type G.

But now, suppose he believes all *and only those* sentences printable by the machine; his belief system then coincides exactly with the set of sentences provable in $\overline{G}$, and since $\overline{G}$ is of type G, then *he* must be of type G. Such a reasoner we will call a *minimal* reasoner of type G. Since the system $\overline{G}$ is self-referentially correct (as we showed in the last chapter), it follows that a minimal reasoner of type G must be wholly accurate in his beliefs. It further follows that any minimal reasoner of type G is both consistent and stable.

We now see that the notion of a consistent, stable reasoner of type G does *not* involve a logical contradiction. A reasoner of type G is

not necessarily consistent (indeed, any *inconsistent* reasoner of even type 1 is also of type G, since he believes everything!), but a *minimal* reasoner of type G is both consistent and stable.

Now let us consider a reasoner who *is* of type G and who is keeping his eye on the output of the machine (and who interprets all the modal sentences egocentrically). Will he necessarily *believe* all these sentences (under the egocentric interpretation)? Well, it is easy to verify that he believes all *axioms* of $\overline{\mathrm{G}}$. (In fact, in Chapter 11 we showed that any reasoner of type 4 knows that he is of type 4; hence he will believe all axioms of $\overline{\mathrm{K}}_4$. A reasoner of type G also believes he is modest, which means that he will believe all sentences of the form $\mathrm{B}(\mathrm{BX}\supset\mathrm{X})\supset\mathrm{BX}$, and so he believes all axioms of $\overline{\mathrm{G}}$.) And since the machine prints nonaxioms only by using the modus ponens and necessitation rules, and the reasoner's beliefs are closed under modus ponens, and he is normal, then he will successively believe every sentence as it gets printed. (If anyone should interrupt the process and ask the reasoner his opinion of the machine, the reasoner will answer: "This machine is truly amazing. Everything it has printed about me so far is true!")

We now see that a reasoner of type G does indeed believe all sentences provable in $\overline{\mathrm{G}}$.

Suppose, next, that a modal sentence X is believed by all reasoners of type G; does it follow that X is actually provable in $\overline{\mathrm{G}}$? The answer is yes, since if X is believed by *all* reasoners of type G, then it must be believed by a *minimal* reasoner of type G, and hence must be provable in $\overline{\mathrm{G}}$. And so we now see that a sentence is provable in $\overline{\mathrm{G}}$ *if and only if* it is believed by all reasoners of type G. Put another way, given any minimal reasoner of type G, he believes those and only those sentences that are believed by *all* reasoners of type G.

Of course, when two different reasoners look at the same modal sentence, they interpret it differently—each one interprets "B" as referring to his *own* beliefs (just as the word "I" has different references when used by different people). And so when we speak of a modal sentence being *believed* by all reasoners of type G, we

mean believed by each one according to his own egocentric interpretation.

Of course everything we have said about the modal system $\overline{G}$ and reasoners of type G also holds for the modal system $\overline{K_4}$ and reasoners of type 4: A sentence is provable in $\overline{K_4}$ if and only if it is believed by all reasoners of type 4. Likewise, a sentence is provable in $\overline{K}$ if and only if it is believed by all reasoners of type 3.

MORE ON MODAL SYSTEMS AND REASONERS

The results of the problems in the remainder of this chapter are not necessary for the understanding of the next two chapters, but are independently interesting.

1

Suppose a reasoner's beliefs are closed under modus ponens and that for any *axiom* X of $\overline{K_4}$, the reasoner believes X and also believes that he believes X. Will he necessarily believe all sentences that are believed by all reasoners of type 4? (Remember, he may not be normal!) The answer is given following Problem 2.

2

Now, substitute $\overline{G}$ in place of K_4. Will the reasoner above necessarily believe all sentences that are believed by all reasoners of type G?

The answer to the above problems is given by a well-known theorem about the modal systems K_4 and G (and which also applies to $\overline{K_4}$ and $\overline{G}$), which we are about to state and whose proof we will sketch.

Let M be any modal system whose only inference rules are modus ponens and necessitation. We let M′ be that modal system whose axioms are the axioms of M together with all formulas BX, where X is an axiom of M, and whose *only* inference rule is modus ponens. It is obvious that everything provable in M′ is provable in M (because for any axiom X of M, BX is also provable in M, and so all axioms of M′ are provable in M), but in general it is not true that everything provable in M is provable in M′. However, we have the following interesting result:

Theorem 1. If all axioms of K_4 are provable in M, then it *is* true that everything provable in M is provable in M′ (and hence the systems M and M′ prove exactly the same formulas).

Can the reader see how to prove Theorem 1?
(*Hint:* First show that if X is provable in M, then BX is provable in M′. Do this by showing that for any proof $X_1, \ldots, X_n$ in M, all of the formulas $BX_1, \ldots, BX_n$ are successively provable in M′.)

More Detailed Proof. Since modus ponens is a rule of M′ and all tautologies are among the axioms of M′, then M′ is of course of type 1. Also the following three conditions hold for M′:

(1) If BX and B(X⊃Y) are provable in M′, so is BY—because B(X⊃Y)⊃(BX⊃BY) is an axiom of M′ and M′ is of type 1.

(2) If BX is provable in M′, so is BBX—because BX⊃BBX is an axiom of M′ and M′ is of type 1.

(3) If X is an axiom of M, then BX is provable in M′—because it is even an axiom of M′.

Now, suppose a sequence $X_1, \ldots, X_n$ of formulas constitutes a proof in M. Each line of the proof either comes from two earlier lines by modus ponens, or from one earlier line by the necessitation rule, or is itself an axiom of M. Using facts (1), (2), and (3) above, it easily

follows that BX_1 must be provable in M′, then that BX_2 is provable in M′, then that BX_3 is provable in M′, and so forth down to BX_n. We leave the verification of this to the reader.

Now we know that if X is provable in M, then BX is provable in M′. Therefore, if X is provable in M′, then BX is provable in M′ (because if X is provable in M′, it is also provable in M). And so we see that M′ is normal (even though the necessitation rule is not initially given for M′). Then, given any proof $X_1, \ldots, X_n$ in M, it is easy to see that each of the lines $X_1, \ldots, X_n$ *itself* can be successively proved in M′. (We leave the verification of this to the reader.)

Corollary. The provable formulas of K_4 and K_4' are the same. The provable formulas of G and G′ are the same.

The following theorem and its corollary can be proved by a similar argument.

Theorem $\overline{1}$. For any *sentential* modal system $\overline{M}$ in which all axioms of $\overline{K_4}$ are provable, and in which modus ponens and necessitation are the only inference rules, the provable sentences of $\overline{M}$ are the same as the provable sentences of $\overline{M}'$.

Corollary. The provable sentences of $\overline{K_4}$ are the same as those of $\overline{K_4}'$. The provable sentences of $\overline{G}$ are the same as those of $\overline{G}'$.

Of course the above corollary gives an affirmative answer to Problems 1 and 2.

We see by virtue of the corollary to Theorem 1 that the modal systems K_4 and G can be alternatively axiomatized using systems whose *only* inference rule is modus ponens.

• 29 •

Some Strange Reasoners!

REASONERS WHO ARE ALMOST OF TYPE G

We shall say that a reasoner is *almost* of type G if he believes all sentences believed by all reasoners of type G (or, what is the same thing, if he believes all sentences provable in the modal system $\overline{G}$) and if his beliefs are closed under modus ponens. What possibly keeps him from being a reasoner of type G is that he may not be normal.

As we will prove, a reasoner who is almost of type G, unlike a reasoner who is of type G, *can* believe that he is consistent without losing his consistency. But then he must suffer from another defect —he cannot be normal!

We will now look more closely into this.

1

Given a consistent reasoner who is almost of type G and who believes he is consistent, find a proposition p such that the reasoner believes p, but can never know that he believes p!

2

Any abnormal reasoner must fail to believe at least one true proposition, because there is a proposition p such that he believes p but fails to believe Bp, yet Bp is true (since he believes p). Therefore he fails to believe the true proposition Bp.

It then follows from the last problem that given any consistent reasoner who is almost of type G and believes he is consistent, there must be at least one true proposition that he fails to believe. More startling is the fact that there must be at least one *false* proposition that he does believe!

What false proposition must he believe?

Exercise 1. State whether the following is true or false: Every nonnormal reasoner of type 1 is consistent.

Exercise 2. State whether the following is true or false: Every nonnormal reasoner of type 1 who believes all axioms of $\overline{K_4}$ must believe at least one false proposition.

REASONERS WHO ARE OF TYPE G*

As we will see, a reasoner who is almost of type G can not only believe in his own consistency without necessarily being inconsistent; he can even believe in his own *accuracy* without necessarily being inconsistent.

By a reasoner of type G* we shall mean a reasoner who is *almost* of type G and who believes all sentences of the form $BX \supset X$ (he believes in his own accuracy). In other words, a reasoner of type G* is a reasoner of type 1 who believes all sentences provable in G and believes all sentences of the form $BX \supset X$.

Such a reasoner must, of course, also believe the sentence $B\bot \supset \bot$, and since he is almost of type G, then what we have proved in the

solution to Problems 1 and 2 also holds good for him. And so we have established Theorem 1.

Theorem 1. For any consistent reasoner of type G*:

(a) He believes that he is consistent, but can never know that he believes he is consistent!

(b) He also believes the false proposition $B\sim B\bot\supset BB\sim B\bot$ (i.e., he wrongly believes: "If I ever believe that I am consistent, then I will believe that I believe that I am consistent.")

We remind the reader that the sentence $B\sim B\bot\supset BB\sim B\bot$ is false for a consistent reasoner of type G*, since $B\sim B\bot$ is true (he believes he is consistent), but $BB\sim B\bot$ is false (he doesn't believe that he believes that he is consistent).

It of course follows from Theorem 1 that *any* reasoner of type G* must have at least one false belief, because if he is consistent, then he does (by Theorem 1), and if he is inconsistent, he certainly does!

Minimal Reasoners of Type G*. By the modal system G* is meant the system whose axioms are all the provable formulas of G together with all formulas of the form $BX\supset X$, and whose only inference rule is modus ponens. We shall let $\overline{G^*}$ be the system G* whose axioms are restricted to sentences—that is, the axioms of $\overline{G^*}$ are all sentences provable in $\overline{G}$, plus all *sentences* of the form $BX\supset X$. The only inference rule of $\overline{G^*}$ is modus ponens. (We can easily show by an argument similar to the one used in Chapter 27 that the provable sentences of $\overline{G^*}$ are the same as the provable *sentences* of G*.) By a *minimal* reasoner of type G* we shall mean a reasoner who believes those and only those sentences provable in $\overline{G^*}$. It is easy to show that all reasoners of type G* must believe all sentences provable in $\overline{G^*}$ ("believe" here falls under the egocentric interpretation, of course). Therefore a reasoner is a minimal reasoner of type G* if and only if he believes those and only those sentences that are believed by all reasoners of type G*.

Since every reasoner of type G* has at least one false belief, so does a minimal reasoner of type G*. It hence follows that there is at least one sentence provable in $\overline{G^*}$ that is false for $\overline{G^*}$ (false, when "B" is interpreted as provability in $\overline{G^*}$). And so we have Theorem 2.

Theorem 2. The modal system $\overline{G^*}$ is *not* self-referentially correct.

In light of Theorem 2, the reader may well wonder how the modal system $\overline{G^*}$ could be of any use. Well, just because there is a provable sentence of $\overline{G^*}$ that is false *for* $\overline{G^*}$ does not mean that there isn't some other interpretation of "B" in which all provable sentences of $\overline{G^*}$ are true. Is there such an interpretation? Yes, there is—and a very important one.

Theorem 3. Every sentence provable in $\overline{G^*}$ is true *for the modal system* $\overline{G}$.

This means that every sentence provable in $\overline{G^*}$ is true *if "B" is interpreted as provability in* $\overline{G}$*, rather than in* $\overline{G^*}$.

3

Why is Theorem 3 correct?

Theorem 4 is an easy corollary of Theorem 3.

Theorem 4. The system $\overline{G^*}$ is consistent.

4

Why is Theorem 4 a corollary of Theorem 3?

Theorem 4, of course, implies that any *minimal* reasoner of type G* is consistent. And so a minimal reasoner of type G* is consistent,

he believes he is consistent, but can never believe that he believes he is consistent (by Theorem 2). Stated alternatively in terms of the modal system $\overline{G^*}$: it is consistent, it can prove its own consistency, but can never prove that it can prove its own consistency! Also, the modal system $\overline{G^*}$ is not normal.

The Completeness of $\overline{G^*}$ for $\overline{G}$. We shall now state a further result whose proof unfortunately goes beyond the scope of this book.

Given two modal systems M_1 and M_2, we have defined M_1 to be *correct* for M_2 if every sentence provable in M_1 is *true* for M_2. Let us say that M_1 is *complete* for M_2 if every sentence that is true for M_2 is actually provable in M_1.

Theorem 3 says that the modal system $\overline{G^*}$ is *correct* for the modal system $\overline{G}$. Well, it also happens to be *complete* for $\overline{G}$—every sentence *true* for $\overline{G}$ is *provable* in $\overline{G^*}$. And so the provable sentences of $\overline{G^*}$ are precisely the sentences that are true for $\overline{G}$. Thus a sentence is provable in $\overline{G^*}$ if and only if it is *true* for all reasoners of type G.

REASONERS OF TYPE Q (QUEER REASONERS)

By a *queer* reasoner—or a reasoner of type Q—we shall mean a reasoner of type G who believes that he is *inconsistent.* Can a queer reasoner be consistent? We will soon see that he can! Of course, every queer reasoner is normal.

By the modal system $\overline{Q}$, we shall mean the modal system $\overline{G}$ with the sentence $B\bot$ added as an axiom. By a *minimal* reasoner of type Q, we mean a reasoner who believes those and only those sentences provable in the modal system $\overline{Q}$—or, what is the same thing, a reasoner who believes all and only those sentences believed by all reasoners of type Q.

Theorem 5. The modal system $\overline{Q}$ is not self-referentially correct, but it *is* consistent.

5

Why is Theorem 5 true?

It of course follows from Theorem 5 that any minimal reasoner of type Q is consistent, although he believes he isn't!

A comparison. It is amusing and instructive to compare minimal reasoners of types G, G*, and Q.

(1) A minimal reasoner of type G is consistent, but can never know it.

(2) A minimal reasoner of type G* is consistent, believes he is consistent, but can never know that he believes he is consistent.

(3) A minimal reasoner of type Q believes he is inconsistent, but he is wrong—he is actually consistent.

SOLUTIONS

1 · One such proposition p is the proposition that the reasoner is consistent!

We are given that he believes $\sim B\bot$ and we are to show that if he is consistent, he cannot believe $B\sim B\bot$. Well, since he believes all sentences provable in G, he certainly believes all tautologies, and since his beliefs are closed under modus ponens, he is certainly of type 1. He believes $\sim B\bot$, so he believes $B\bot\supset\bot$. If he believed $B\sim B\bot$, he would believe $B(B\bot\supset\bot)$. However, he does believe $B(B\bot\supset\bot)\supset B\bot$ (because he believes all sentences provable in G). And so he would then believe $B(B\bot\supset\bot)$ and believe $B(B\bot\supset\bot)\supset B\bot$, hence he would believe $B\bot$. But since he believes $\sim B\bot$, he would be inconsistent.

This proves that if he believes $B\sim B\bot$, he would be inconsistent. But we are given that he is consistent, hence he can never believe $B\sim B\bot$ (even though $B\sim B\bot$ is true).

2 · $B\sim B\bot$ is true, but since he doesn't believe it, then $BB\sim B\bot$ is false —hence $B\sim B\bot \supset BB\sim B\bot$ is false. But he must believe this sentence (because it is of the form $BX\supset BBX$ where $X=\sim B\bot$), hence it is an axiom of G. And so he believes the false sentence $B\sim B\bot \supset BB\sim B\bot$. (He wrongly believes: "If I should believe I am consistent, then I would believe that I believe that I am consistent." This belief is wrong, since in fact he *does* believe he is consistent, but *doesn't*—and never will—believe that he believes he is consistent.)

Incidentally, by the same argument, *any* nonnormal reasoner who believes all sentences provable in K_4 must have at least one false belief. There is some p such that he believes p but doesn't believe Bp, so $Bp\supset BBp$ is false (for such a reasoner), yet it is an axiom of K_4, and the reasoner therefore believes it. This answers Exercise 2.

Exercise 1. It is true! If he were inconsistent and of type 1, he would believe *all* propositions, hence there would be no proposition p such that he believes p and doesn't believe Bp (because he believes both p and Bp, since he believes everything). Therefore every inconsistent reasoner of type 1 must be normal—or, put another way, every nonnormal reasoner of type 1 is consistent.

3 · We proved in the last chapter that G is self-referentially correct, hence:

(1) All sentences provable in $\overline{G}$ are true for $\overline{G}$.

Also:

(2) All sentences of the form $BX\supset X$ are true for $\overline{G}$.

The reason for (2) is that if BX is true for $\overline{G}$, then X is provable in $\overline{G}$ (that's what it means for BX to be true for $\overline{G}$), and X must therefore be true for $\overline{G}$ (since $\overline{G}$ is self-referentially correct). And so, if BX is true for $\overline{G}$, so is X—which means that $BX\supset X$ is true for $\overline{G}$.

By virtue of (1) and (2), every *axiom* of $\overline{G^*}$ is true for $\overline{G}$. Since the only inference rule of $\overline{G^*}$ is modus ponens, and since the set of

sentences that is *true* for $\overline{G}$ is closed under modus ponens (if X and X⊃Y are true for $\overline{G}$, then obviously Y is true for $\overline{G}$), it follows that every sentence provable in $\overline{G^*}$ must be true for $\overline{G}$. Thus the system $\overline{G^*}$ is *correct* for $\overline{G}$.

4 · Since $\overline{G^*}$ is correct for $\overline{G}$, then if ⊥ were provable in $\overline{G^*}$, it would be true for $\overline{G}$, which is absurd. Therefore ⊥ is not provable in $\overline{G^*}$, and so $\overline{G^*}$ is consistent (even though it is not self-referentially correct).

5 · The system $\overline{Q}$ is of type G, and therefore by (c) of Lemma B, Chapter 27 (page 233), all axioms of $\overline{G}$ are true for $\overline{Q}$.

Now, let us momentarily assume that the sentence B⊥ is true for $\overline{Q}$. We then get the following contradiction: If B⊥ is true for $\overline{Q}$, then since all the other axioms of $\overline{Q}$ (i.e., the axioms of $\overline{G}$) are true for $\overline{Q}$, we would have *all* the axioms of $\overline{Q}$ being true for $\overline{Q}$. Then by Corollary A_1 of Lemma A, Chapter 27 (page 231), the system $\overline{Q}$ would be self-referentially correct. And so, if B⊥ is true for $\overline{Q}$, then $\overline{Q}$ is self-referentially correct. On the other hand, to say that B⊥ is true for $\overline{Q}$ is to say that ⊥ is provable in $\overline{Q}$, and since ⊥ is obviously false for $\overline{Q}$, this would mean that $\overline{Q}$ is *not* self-referentially correct. It is therefore contradictory to assume that B⊥ is true for $\overline{Q}$. Hence B⊥ is false for $\overline{Q}$, which means that ⊥ is *not* provable in $\overline{Q}$, and so $\overline{Q}$ must be consistent! But also B⊥ is an axiom of $\overline{Q}$, hence of course provable in $\overline{Q}$, and since it is false for $\overline{Q}$, then $\overline{Q}$ is not self-referentially correct. And so we see that $\overline{Q}$ is consistent but not self-referentially correct.

• 30 •

In Retrospect

WE BEGAN this study with introspective reasoners and have wound up in the labyrinths of modal logic. Let us summarize some of the main things we have learned on the journey.

1. An accurate Gödelian system of type 1 cannot prove its own accuracy—i.e., it cannot prove all propositions of the form $BX \supset X$.

2. Any Gödelian system of type 1 that can prove its own accuracy is not only inaccurate, but peculiar—i.e., there must be a proposition p such that p and $\sim Bp$ are both provable.

3. Any Gödelian system of type 1* which can prove its own nonpeculiarity is peculiar.

4. (After Gödel's First Incompleteness Theorem.) Any normal, stable, consistent Gödelian system of type 1 must be incomplete. More specifically, if S is a normal system of type 1 and p is a proposition such that $p \equiv \sim Bp$ is provable in S, then:

(a) If S is consistent, p is not provable in S.

(b) If S is consistent and stable, then $\sim p$ is also not provable in S.

5. (After Gödel's Second Theorem.) No consistent Gödelian system of type 4 can prove its own consistency.

6. A Gödelian system of type 4 can even prove that if it is consistent, then it cannot prove its own consistency—i.e., it can prove the proposition $\sim B\bot \supset \sim B(\sim B\bot)$.

7. (After Löb.) If S is a reflexive system of type 4, then for any

proposition p of the system, if $Bp \supset p$ is provable in the system, so is p.

8. A system of type 4 is reflexive if and only if it is of type G.

9. A system of type 4 is Löbian if and only if it is of type G.

10. (After Kripke, de Jongh, Sambin.) Any system of type 3 in which all propositions of the form $B(BX \supset X) \supset BX$ are provable must be of type 4 (and hence of type G).

11. A consistent system of type G cannot prove any proposition of the form $\sim BX$—in particular, it cannot prove its own consistency.

12. A consistent and stable system of type G can neither prove its own consistency nor its own inconsistency.

13. (Semantical Soundness Theorems.) For any modal formula X, if X is provable in K, then it holds in all Kripke models; if it is provable in K_4, it holds in all transitive models; and if it is provable in G, then it holds in all transitive terminal models.

14. There do exist consistent stable systems of type G—for example, the machines of Fergusson and Craig, and the modal system $\overline{G}$. These systems, though consistent, cannot prove their own consistency.

15. The modal systems $\overline{K}$, $\overline{K_4}$, and $\overline{G}$ are not only consistent and stable, but are self-referentially correct. The same goes for the systems K, K_4, and G.

16. Neither of the systems $\overline{G^*}$ or $\overline{Q}$ is self-referentially correct, but both of them are consistent. The system $\overline{Q}$ is normal, but the system $\overline{G^*}$ is not.

17. (a) A minimal reasoner of type G is consistent, but can never know it.

(b) A minimal reasoner of type G^* is consistent and believes that he is consistent, but can never know that he believes he is consistent.

(c) A minimal reasoner of type Q believes he is inconsistent, but is really consistent.

This seems like a good stopping place. There are many more fascinating things about the modal system G. By far the best general

reference currently in existence is Boolos's *The Unprovability of Consistency,* which I heartily recommend as a follow-up to this book. To whet the reader's appetite, let me give a sample of a beautiful result—a fixed-point theorem—whose proof can be found there.

Consider a modal formula with only one propositional variable, say the letter p. Let us denote such a formula as A(p). For any modal sentence S, by A(S) we mean the result of substituting S for every occurrence of p in A(p). For example, if A(p) is the formula p⊃Bp, then A(S) is the sentence BS⊃S. A sentence S is called a *fixed point* of A(p) if the sentence S≡A(S) is provable in G. In Problem 4 of Chapter 19, we asked the reader to find a proposition p such that any reasoner of type G will believe p≡~Bp, and we found ~B⊥ to be a solution. Thus every reasoner of type G will believe ~B⊥≡~B~B⊥, hence this sentence is provable in G. This means that ~B⊥ is a fixed point of the formula ~Bp. The formula B~p also has a fixed point—namely, B⊥, as we found in the solution to Problem 5, Chapter 19. The reader might find it a profitable exercise to try to show that the formula Bp⊃B⊥ has a fixed point—namely, BB⊥⊃B⊥.

Not every formula A(p) has a fixed point; for example, the formula ~p doesn't (otherwise the system G would be inconsistent, which we know is not so). Now, a formula A(p) is called *modalized* in p if every occurrence of p in A(p) lies in the part of A(p) of the form BX, where X is a formula. (Examples: Bp⊃BBp is modalized in p; Bp⊃p is not, but B(Bp⊃p) is.) The logicians Claudio Bernardi and C. Smorynski have independently proved that any formula A(p) that *is* modalized in p *does* have a fixed point S—moreover, the formula B(p≡A(p))⊃B(p≡S) is provable in G. This result is known as the Bernardi-Smorynski Fixed-Point Theorem.

Fixed points are remarkable things. By virtue of the self-referential correctness of the system G, any fixed point S of a formula A(p) is not only provable in G if and only if A(S) is provable, but is also *true* (for G) if and only if A(S) is true—because the provability of

$S \equiv A(S)$ implies its truth. Let us say that a formula A(p) *applies* to a sentence S if A(S) is true (for G). A fixed point of a formula can then be thought of as a sentence that asserts that the formula applies to that very sentence.

More generally, consider a formula A(p,q) having no propositional variables other than p and q. For any formula X, by A(X,q) is meant the result of substituting X for p in A(p,q). By a fixed point of A(p,q) is meant a formula H having no variables other than q, such that the formula $H \equiv A(H,q)$ is provable in G. The logicians D. H. J. de Jongh and Giovanni Sambin have proved that if A(p,q) is modalized in p (but not necessarily in q), then A(p,q) has a fixed point H; moreover, the formula $B(p \equiv A(p,q)) \supset B(p \equiv H)$ is provable in G.

Examples are already familiar from Chapter 19. Bq is a fixed point of $B(p \supset q)$, by Problem 6, and $Bq \supset q$ is a fixed point of $Bp \supset q$, by Problem 7. Indeed, reflexivity is equivalent to there being a fixed point for $Bp \supset q$. The remarkable thing is that for a modal system of type 4, the existence of a fixed point for the one formula $Bp \supset q$ is enough to guarantee fixed points for *all* formulas A(p,q) that are modalized in p. A proof of this can also be found in Boolos.

CONCLUDING REMARKS

I hope I have given the reader some feeling for why these modal systems are so interesting. We have seen that they can be interpreted both internally (self-referentially) and externally (as applying to provability in other mathematical systems), as well as to reasoning processes—both for naturally intelligent beings (some humans and other animals) and for artificially intelligent mechanisms (such as computers). What applications this may have in the field of psychology is something that might be worth further investigation.

It is a happy turn of fate that the field of modal logic, which historically arose out of purely philosophical interests, should have turned out to be so important today in proof theory and computer

science—this by virtue of the theorems of Gödel and Löb and of the work of those who have subsequently looked at proof theory from a modal-theoretical viewpoint. And now even those philosophers who in the past have taken a dim view of the significance of modal logic are forced to realize its mathematical importance.

The past philosophical opposition to modal logic has been grounded roughly in three quite different (and incompatible) beliefs: First, there are those who believe that everything true is necessarily true, and hence that there is no difference between truth and necessary truth. Second, there are those who believe that nothing is necessarily true, and hence that for any proposition p, the proposition Np (p is necessarily true) is simply false! And third, there are those who claim that the words "necessarily true" convey no meaning whatsoever. And so each of these philosophical types has rejected modal logic on his own grounds. Indeed, one well-known philosopher is reputed to have suggested that modern modal logic was conceived in sin. To which Boolos has aptly replied: "If modern modal logic was conceived in sin, then it has been redeemed through Gödliness."