• *Part IV* •

LET'S BE CAREFUL!

• 9 •

Paradoxical?

WE NOW have the background to embark on our journey to Gödel's consistency predicament, which we will reach in Chapter 12, encountering many interesting problems along the way. Let us start with a problem closely related to one of the variants of the Surprise Examination Paradox in Chapter 2.

We are back on the Island of Knights and Knaves, where the following three propositions hold: (1) knights make only true statements; (2) knaves make only false ones; (3) every inhabitant is either a knight or a knave. These three propositions will be collectively referred to as the "rules of the island."

We recall that no inhabitant can claim that he is not a knight, since no knight would make the false statement that he isn't a knight and no knave would make the true statement that he isn't a knight.

Now suppose a logician visits the island and meets a native who makes the following statement to him: *"You will never know that I am a knight."*

Do we get a paradox? Let us see. The logician starts reasoning as follows: "Suppose he is a knave. Then his statement is false, which means that at some time I *will* know that he is a knight, but I can't *know* that he is a knight unless he really is one. So, if he is a knave, it follows that he must be a knight, which is a contradiction. Therefore he can't be a knave; he must be a knight."

So far, so good—there is as yet no contradiction. But then he continues reasoning: "Now I know that he is a knight, although he

said that I never would. Hence his statement was false, which means that he must be a knave. Paradox!"

Question. Is this a genuine paradox?

Discussion. This problem bears a good deal of analysis! To begin with, the paradox (if it really is one) is what would be called a *pragmatic* paradox rather than a purely logical one, since it involves not only logical notions such as truth and falsity, but pragmatic notions such as *knowing.* To emphasize the pragmatic nature of the question, is it not possible that whether or not a paradox arises might depend on the person to whom the statement is made? It certainly is! As an extreme example, the native could surely say it to a dead person, and no paradox would arise. (He points to the corpse and says: "You will never know that I'm a knight." Well, he is certainly right; the corpse will indeed never know that he is a knight, and so the native is in fact a knight—since what he said is true. But since the corpse will never know it, no contradiction arises.) To take a less extreme example, the native might say this to a person who is alive but deaf and hence does not hear the statement; again, no paradox would arise.

So we must assume that the visitor to the island was alive and heard the statement, but this is still not enough. We must assume a certain *reasoning* ability on the visitor's part, because if the visitor has no reasoning ability, he would not go through the argument I have given. (The native would say: "You will never know that I'm a knight." The visitor might then say: "That's interesting," walk away, and never think about the matter again. Hence no paradox would arise.) And so we must make explicit what reasoning abilities the logician has.

We will define an individual to be a *reasoner of type 1* if he thoroughly understands propositional logic—that is, if the following two conditions hold:

(1) He believes all tautologies.

(2) For any propositions X and Y, if he believes X and believes X⊃Y, then he believes Y.

In the terminology of Chapter 8, the set of beliefs of a reasoner of type 1 is logically closed. It then follows by Principle L of Chapter 8, that given any finite set S of propositions that he believes, he must believe all logical consequences of S as well.

We shall now make the assumption that the visitor to the island is a reasoner of type 1. Of course this assumption is highly idealized, since there are *infinitely* many tautologies, hence our assumption implies something like immortality on the reasoner's part. However, little things like that don't bother us in the timeless realm of mathematics. We simply imagine the reasoner so programmed that (1) sooner or later he will believe every tautology; (2) if he ever believes p and ever believes p⊃q, then sooner or later he will believe q. It then follows by Principle L of Chapter 8 that given any finite set S of propositions, if the reasoner believes all the propositions in S, then for any proposition Y that is a logical consequence of S, the reasoner will sooner or later believe Y.

We still need further assumptions. For one thing we must assume a certain *self-consciousness* on the reasoner's part; specifically, we must assume that if the reasoner ever knows something, then he knows that he knows it (otherwise he could never have said, "Now I know that he's a knight," and my argument wouldn't go through).

Considering the problem with all these assumptions, does a genuine paradox now arise? Still not, for although I told you that the rules of the island hold (every inhabitant is either a knight or a knave; knights make only true statements; knaves make only false statements), we must make the additional assumption that the reasoner *believes* that the rules of the island hold. Indeed, it is perfectly reasonable that the reasoner might believe this at the outset, but after finding himself in a contradiction following the argument I gave, he would have rational grounds for *doubting* the rules of the

island. (I imagine you and I would do just that after finding ourselves in such a predicament!) Well, to make the problem really interesting, let us make as our final assumption that the reasoner believes *and continues to believe* the rules of the island.

1.

Now we run into an interesting problem. Under the additional assumptions we have made, the reasoner's argument that the native is a knight and that the native is a knave seems perfectly valid. Yet the native can't be both a knight and a knave! So what is wrong with the reasoner's argument?

Solution. I phrased the above problem in a very misleading way. (Occasionally I feel like being a bit sneaky!) It is not the reasoner's argument that was wrong; it's that the situation I described could never have arisen. If a reasoner of type 1 comes to a knight-knave island and believes the rules of the island (and hears whatever statements are made to him), then it is logically impossible that any native will say to him, "You will never know that I am a knight."

To prove this, we don't need one of the assumptions we have made—namely, that if the reasoner knows something, then he knows that he knows it. Even without this assumption, we can get a contradiction as follows: The reasoner reasons, "Suppose he is a knave. Then his statement is false, which means that I will know that he is a knight, which implies that he really is a knight. Therefore the assumption that he is a knave leads to a contradiction, so he must be a knight."

Without going any further in the logician's reasoning process, we can derive a contradiction. The logician has so far reasoned correctly and has come to the conclusion that the native is a knight. Since he has reasoned correctly, then the native really is a knight, and so the reasoner *knows* that the native is a knight. However, the native said he would never know that, hence the native must be a knave.

Therefore the native is both a knight and a knave, which is a contradiction.

Instead of using the word "know," we could just as well have said "correctly believe," and our argument would still go through. We say that an individual *correctly believes* a proposition p if he believes p and p is true.

We have now proved Theorem I.

Theorem I. Given a knight-knave island and a reasoner of type 1 who believes the rules of the island (and who hears any statement addressed to him), it is logically impossible that any native can say to him, "You will never correctly believe that I am a knight."

Discussion. Some critical readers might object to the proof I have given of the above theorem on the grounds that I have credited the reasoner with more abilities than I have explicitly ascribed to him —namely, that the reasoner makes assumptions and subsequently discharges them. In the specific case in hand, the reasoner started out: "Suppose he is a knave. Then —— ." Well, this is really only a matter of convenience, not necessity. I could have given the reasoner's argument in the following, more direct form: "If he is a knave, then his statement is false. If his statement is false, then I'll correctly believe he is a knight. If I correctly believe he is a knight, then he is a knight. Putting these three facts together, if he is a knave, then he is a knight. From this last fact it logically follows that he is a knight."

It is a common practice in logic to prove a proposition of the form $p \supset q$ (if p then q) by supposing that p is true and then trying to derive q. If this can be done, then the proposition $p \supset q$ is established. In other words, if assuming p as a premise leads to q as a conclusion, then the proposition $p \supset q$ has been proved. (This technique is part of what is known as *natural deduction.*) The whole point is that anything that can be proved using this device can also be proved without it. (There is a well-known theorem in logic to this effect; it

is called the *deduction theorem.*) And so we shall allow our "reasoners" to use natural deduction; a reasoner of type 1 who does this cannot prove any more facts than he can without using natural deduction, but the proofs using natural deduction are usually shorter and easier to follow. Therefore we shall continue to let our reasoners use natural deduction.

2 · A Dual Problem

We continue to assume that the reasoner is of type 1, that he believes the rules of the island, and that he hears all remarks addressed to him.

Suppose the native, instead of saying, "You will never correctly believe I'm a knight," says, "You will correctly believe I'm a knave."

Do we then get a contradiction? (The reader should try solving this before reading the solution.)

Solution. The reasoner reasons as follows: "Suppose he is a knight. Then his statement is true, which means that I will correctly believe he is a knave, which in turn implies that he is a knave. Hence the assumption that he is a knight leads to a contradiction, therefore he must be a knave."

At this point the reasoner believes the native is a knave, and he has reasoned correctly, hence the native is a knave. On the other hand, since the reasoner correctly believes that the native is a knave, the native's statement was true, which makes him a knight. So we do indeed get a contradiction.

SOME RELATED PROBLEMS

Let us now leave the Island of Knights and Knaves for a while and consider a problem related to the paradox of Chapter 3. A student asks his theology professor: "Does God really exist?" The professor

gives the following curious answer: "God exists if and only if you don't correctly believe that He does."

3

Suppose that the student is a reasoner of type 1 and that the professor's statement is true and that the student believes the statement. Do we then get a paradox?

Solution. Yes, we do! To begin with, even forgetting that the student is a reasoner of type 1 and that he believes the professor's statement, it follows that God must exist, because if God didn't exist, then the student *would* correctly believe that God exists, but no one can *correctly* believe a false proposition. Therefore God must really exist (assuming that the professor's statement is true).

Now, the student, being a reasoner of type 1, knows propositional logic as well as you or I, hence he also is able to reason that if the professor's statement is true, then God must exist. But he also believes the professor's statement; therefore he must believe that God exists. And since we have proved that God exists (under the three assumptions of the problem), then the student *correctly* believes that God exists. But God exists if and only if the student *doesn't* correctly believe that God exists. From this it follows that God doesn't exist if and only if the student *does* correctly believe that God exists. (For any proposition p and q, the proposition $p \equiv \sim q$ is logically equivalent to the proposition $\sim p \equiv q$.) Since God doesn't exist if and only if the student correctly believes that God exists, and the student does correctly believe that God exists, then it follows that God doesn't exist. Thus the three assumptions of the problem lead to the paradox that God does exist and doesn't exist.

Of course the same paradox would arise if the professor had instead said: "God exists if and only if you correctly believe that He doesn't exist." We leave the proof of this as an exercise for the reader.

It is now important for us to realize that the above paradox is essentially the same as that of Problem 1 concerning the knight-knave island, although they may appear different. The seeming differences are: (1) In the above paradox, the professor made an "if and only if" statement, whereas in Problem 1, the native did not; he said outright that the reasoner would never correctly believe that the native is a knight. (2) In the above paradox, the student believes the professor, whereas in Problem 1, the reasoner has no initial belief that the native is a knight. However, these two differences in a sense cancel each other out, as we will now see. The key to this is the translation device in Chapter 7.

In virtually all the problems that follow, we will be dealing with only two individuals—the native of the island who makes the statement, and the reasoner who hears the statement. We will consistently use the letter k for the proposition that the native in question is a knight. Then, as we saw in Chapter 7, whenever the native asserts a proposition q, the proposition $k \equiv q$ is true. Now, the reasoner *believes* the rules of the island (he believes that knights make true statements and knaves make false ones), and we are assuming that he hears any statement made to him. Therefore, whenever the native asserts a proposition q to the reasoner, the reasoner *believes* the proposition $k \equiv q$. Indeed, from now on, when we say that the rules of the island hold, we need mean no more than that for any proposition q, if the native asserts q, then the proposition $k \equiv q$ is true. And when we say that the reasoner believes the rules of the island (and hears all statements made to him), we need mean no more than that for any proposition q, if the native asserts q to the reasoner, then the reasoner believes the proposition $k \equiv q$.

For any proposition p, we let Bp be the proposition that the reasoner believes (or will believe) p. And we let Cp be the proposition p&Bp. We read Cp as "The reasoner *correctly* believes p."

Now, in Problem 1, the native asserted the proposition $\sim Ck$ ("You will never correctly believe that I am a knight"). Since the rules of the island hold, then the proposition $k \equiv \sim Ck$ is true. Since

the reasoner believes the rules of the island, then he believes the proposition k≡~Ck. These two facts turn out to be logically incompatible, hence a paradox arises.

In Problem 3, let g be the proposition that God exists. The professor asserted outright the proposition g≡~Cg. Under the assumption that the professor makes only true statements, the proposition g≡~Cg must be true. Since the student believed the professor, then he believed the proposition g≡~Cg. Again, the truth of g≡~Cg turned out to be logically incompatible with the student believing g≡~Cg (since the student is a reasoner of type 1).

We now see exactly what the two paradoxes have in common; in both cases we have a proposition p (which is k, for Problem 1, and g for Problem 3) such that p≡~Cp is both true and believed by the reasoner—in other words, it is *correctly* believed by the reasoner, and this is logically impossible if the reasoner is of type 1. Thus both paradoxes (or rather their resolutions that the given conditions are logically incompatible) are special cases of the following theorem.

Theorem A. There is no proposition p such that a reasoner of type 1 can correctly believe the proposition p≡~Cp. In other words, there is no proposition such that a reasoner of type 1 can correctly believe: "The proposition is true if and only if I don't correctly believe that it is true."

The proof of Theorem A is little more than a repetition of the two special cases already considered, but it may help to consider it in a more general setting and to point out some of its interesting features.

To begin with, for any propositions p and q, the proposition (p≡~(p&q))⊃p is a tautology (as the reader can verify). In particular, the proposition (p≡~(p&Bp))⊃p is a tautology. We are letting Cp be the proposition p&Bp, and so (p≡~Cp)⊃p is a tautology. Suppose now that a reasoner of type 1 correctly believes p≡~Cp, we then get the following contradiction: Since the reasoner *correctly* believes p≡~Cp, then p≡~Cp must be true. Also (p≡~Cp)⊃p is

true (it is a tautology), and so p must be true. Now since the reasoner is of type 1, he *believes* the tautology $(p \equiv \sim Cp) \supset p$ and he also believes $p \equiv \sim Cp$ (by assumption), and since he is of type 1, he will then believe p. And so p is true, and he believes p, so he correctly believes p. Thus Cp is true, hence $\sim Cp$ is false. But since p is true and $\sim Cp$ is false, it cannot be that $p \equiv \sim Cp$ (since a true proposition cannot be equivalent to a false proposition), and so we get a contradiction from the assumption that a reasoner of type 1 correctly believes $p \equiv \sim Cp$.

There is also the following "dual" of Theorem A whose proof we leave to the reader.

Theorem A°. There is no proposition p such that a reasoner of type 1 can correctly believe $p \equiv C(\sim p)$.

Exercise 1. Prove Theorem A°.

Exercise 2. Suppose we have a perfectly arbitrary operation B which assigns to every proposition p a certain proposition Bp. (What the proposition Bp is needn't be specified; in this chapter, we have let Bp be the proposition that the reasoner *believes* p; in a later chapter, in which we will be discussing mathematical systems rather than reasoners, Bp will be the proposition that p is *provable* in the system. But for now, Bp will be unspecified.) We let Cp be the proposition $(p \& Bp)$.

(a) Show that one can derive a logical contradiction from the following assumptions:

(i) All propositions of the form BX, where X is a tautology.

(ii) All propositions of the form $(BX \& B(X \supset Y)) \supset BY$.

(iii) Some proposition of the form $C(p \equiv \sim Cp)$.

(b) Show that we also get a logical contradiction if we replace (iii) with "Some proposition of the form $C(p \equiv C \sim p)$."

(c) Why is Theorem A a special case of (a) above? Why is Theorem A° a special case of (b)?

• 10 •

The Problem Deepens

CONCEITED REASONERS

We are back to the Island of Knights and Knaves. Suppose, now, that the native, instead of saying: "You will never correctly believe I'm a knight," makes the following statement: *"You will never believe that I am a knight."*

The native has left out the word "correctly," and as a result things will get far more interesting. We continue to assume that the one addressed is a reasoner of type 1 and that he believes the rules of the island (and also that he has heard the statement) and that the rules of the island really hold. And now we shall make the further assumption that the reasoner is completely accurate in his judgments; he doesn't believe any proposition that is false. Do we still get a paradox?

Well, suppose the native is a knave. Then his statement is false, which means that the reasoner will believe he is a knight. And since the reasoner is accurate in his judgments, then the native really is a knight. Thus the assumption that the native is a knave leads to a contradiction, so the native must be a knight.

Now, the reasoner is of type 1 and knows as much logic as you and I. What is to prevent him from going through the same reasoning process that we just went through and coming to the same

conclusion—namely, that the native must be a knight? Therefore the reasoner will *believe* that the native is a knight, which makes the native's statement false, hence the native must be a knave. But we have already proved that the native is a knight. Paradox!

1

The above argument is fallacious! Can the reader spot the fallacy? (Hint: Despite the fact that the reasoner knows propositional logic as well as you and I, there is something we know that the reasoner doesn't know. What is it?)

Solution. I told you that the reasoner is always accurate; I never said that he *knew* he was accurate! If he knew he was accurate (which in fact he can't know), we would get a paradox. You see, part of our proof that the native is a knight used the assumption that the reasoner is always accurate; if the reasoner made the same assumption, then he could likewise prove that the native is a knight, thus making the native a knave.

Let us now retract the assumption that the reasoner is always accurate in his judgments, but let us suppose that the reasoner *believes* that he is always accurate. Thus for any proposition p, the reasoner believes that if he should ever believe p, then p must be true. Such a reasoner we will call a *conceited* reasoner. Thus a conceited reasoner is one who believes that he is incapable of believing any false proposition.

And so we retract the assumption that the reasoner is always accurate and replace it with the assumption that the reasoner *believes* that he is always accurate. Do we then get a paradox? No, we don't; we instead have the following more interesting result.

Theorem 1. Suppose a native of a knight-knave island says to a reasoner of type 1: "You will never believe I'm a knight." Then if

the reasoner believes himself always accurate, he will lapse into an inaccuracy—i.e., he will sooner or later believe something false.

2

Prove Theorem 1.

Solution. The reasoner reasons: "Suppose he is a knave. Then his statement is false, which means that I will believe he is a knight. But if I ever believe he is a knight, he must really be one, because I am not capable of making mistakes [sic!]. So if he is a knave, he is a knight, which is not possible. Therefore he is not a knave; he is a knight."

At this point, the reasoner believes the native is a knight. Since the native said that the reasoner would never believe that, then the native is in fact a knave. So the reasoner now has the *false* belief that the native is a knight.

The interesting thing is that if the reasoner had been more modest and had not assumed his own infallibility, he would never have been driven into the inaccuracy of believing the native a knight. The reasoner has been justly punished for his conceit!

PECULIAR REASONERS

Let us say that a reasoner is *accurate* with respect to a given proposition p if the reasoner's believing p implies that p is true; in other words, if it is either not the case that he believes p, or it is the case that he believes p and p is true. We will say that the reasoner is inaccurate with respect to p if he believes p and p is false.

It is a noteworthy fact about the last problem that the reasoner was inaccurate with respect to the very proposition about which he believed himself to be accurate—namely, the proposition that the native is a knight. By believing that he was accurate with respect to

that proposition, he finally came to believe that the native was a knight, thus making the native a knave. Of course our proof that the native is a knave rested on our assumption that the rules of the island held. Suppose we retract this assumption but continue to assume that the reasoner *believes* that the rules of the island hold; does it still follow that the reasoner will believe some false proposition? Of course the reasoner will still believe that the native is a knight, although the native said he never would; but if the rules of the island don't hold, then the native is not necessarily a knave. However, even though our proof of Theorem 1 fails if we retract the assumption that the rules of the island hold, we have the following more startling proposition:

Theorem 2. Suppose an inhabitant of the island says to a reasoner of type 1: "You will never believe that I'm a knight." Suppose the reasoner *believes* that the rules of the island hold. Then regardless of whether the rules really hold or not, if the reasoner is conceited, he will come to believe some false proposition.

3

Under the assumption of Theorem 2, what false proposition will the reasoner believe?

Solution. By the same method of proof we followed for Theorem 1, the reasoner will come to believe that the native is a knight. This belief is not necessarily false (since the rules of the island don't necessarily hold), but then the reasoner continues: "Since he is a knight and he said I would never believe he is a knight, then what he said must be true—namely, that I don't believe (i.e., it's not the case that I believe) he's a knight. So I don't believe he is a knight."

At this point, the reasoner believes that the native is a knight and also believes that he doesn't believe that the native is a knight. So

he has the false belief that he doesn't believe the native is a knight (it's false, since he *does* believe that the native is a knight!).

The conclusion of the above problem is really quite weird. The reasoner believes that the native is a knight and also believes that he doesn't believe that the native is a knight. Now, this does not involve a logical inconsistency on the part of the reasoner, though it certainly does involve a psychological peculiarity. We shall call a reasoner *peculiar* if there is some proposition p such that he believes p and also believes that he doesn't believe p. This condition of course implies that the reasoner is inaccurate (because he believes the false proposition that he doesn't believe p). And so a peculiar reasoner is automatically inaccurate, but not necessarily inconsistent.

Let us say that a reasoner is *peculiar with respect to a given proposition p* if he believes p and also believes that he doesn't believe p. A reasoner is then peculiar if and only if there is at least one proposition p with respect to which he is peculiar. If a reasoner is peculiar with respect to p, then he is inaccurate, not necessarily with respect to p, but with respect to Bp.

We now see that even if we remove the assumption that the rules of the island actually hold, if the reasoner believes that they hold and he is of type 1 and a native tells him that he will never believe the native is a knight, then if the reasoner believes that he is accurate with respect to the proposition that the native is a knight, his belief forces him to be inaccurate with respect to the proposition that he *believes* the native is a knight. If the rules of the island do actually hold, then the reasoner will also be inaccurate with respect to the proposition that the native is a knight—i.e., he will believe that the native is a knight, whereas the native is really a knave.

Of course this same problem can be formulated in the context of the student and his theology professor. Suppose the student is a reasoner of type 1 and his professor says to him: "God exists if and

only if you will never believe that He does." Suppose also that the student believes that if he ever believes that God exists, then God does exist. This will mean: (1) If the student believes the professor, then he will wind up believing that God exists and also believing that he doesn't believe that God exists. (2) If the professor's statement is also true, then God doesn't exist, but the student will believe that God does exist.

Both versions of this problem are special cases of the following theorem.

Theorem A. Suppose a reasoner is of type 1 and that there is a proposition p such that he believes the proposition $p \equiv \sim Bp$ and also believes that $Bp \supset p$. Then it follows that:

(a) He will believe p and also believe that he doesn't believe p (he will be peculiar with respect to p).

(b) If also $p \equiv \sim Bp$ is true, then p is false, but he will believe p.

Proof. (a) He believes $p \equiv \sim Bp$, hence he believes $Bp \supset \sim p$ (which is a logical consequence of $p \equiv \sim Bp$). He also believes $Bp \supset p$ (by hypothesis), hence he must believe $\sim Bp$ (which is a logical consequence of $Bp \supset \sim p$ and $Bp \supset p$). But he also believes $p \equiv \sim Bp$, hence he must believe p (which is a logical consequence of $\sim Bp$ and $p \equiv \sim Bp$). And so he believes p and he also believes $\sim Bp$ (he believes that he doesn't believe p!).

(b) Suppose also that $p \equiv \sim Bp$ is true. Since $\sim Bp$ is false (he *does* believe p!), then p, being equivalent to the false proposition $\sim Bp$, is also false. Therefore he believes p, but p is false.

Exercise. Suppose a native says to a conceited reasoner of type 1: "You will believe that I am a knave." Prove: (a) If the reasoner believes that the rules of the island hold, then he will believe that the native is a knave and also that he doesn't believe that the native is a knave. (b) If also the rules of the island really hold, then the native is in fact a knight.

REASONERS OF TYPE 1*

By a reasoner of *type 1**, we shall mean a reasoner of type 1 with the added property that for any propositions p and q, if he ever believes the proposition p⊃q, then he will believe that if he ever believes p then he will also believe q. In symbols, if he ever believes p⊃q, then he will also believe Bp⊃Bq.

Let us note that if a reasoner of type 1 does believe p⊃q, then it is true that if he ever believes p, then he will believe q—i.e., Bp⊃Bq is a true proposition (if he believes p⊃q). What a reasoner of type 1* has that a reasoner of just type 1 doesn't have is that if he believes p⊃q, then not only is the proposition Bp⊃Bq a true one, but he correctly *believes* Bp⊃Bq. Thus a reasoner of type 1* has a shade more "self-awareness" than a reasoner who is only type 1.

We continue to assume that the reasoner, who is now of type 1*, believes the rules of the island and hears all statements made to him, and so whenever the native asserts a proposition p, the reasoner believes the proposition k≡p, where k is the proposition that the native is a knight.

The following fact will be quite crucial:

Lemma 1.† Suppose the native asserts a statement to a reasoner of type 1*. Then the reasoner will believe that if he ever believes that the native is a knight, he will also believe what the native said.

Problem 4. How is this lemma proved?

Solution. (The proof is really quite simple!) Suppose the native asserts the proposition p to the reasoner. Then the reasoner believes

†A lemma is a proposition proved not so much for its own sake as for help in proving subsequent theorems. You might say that a lemma is a proposition that is not "dignified enough" to be called a theorem.

the proposition k≡p. Then he also believes k⊃p, because k⊃p is a logical consequence of k≡p. Then, since he is of type 1*, he will believe Bk⊃Bp.

I have called a reasoner "conceited" if he believes in his own infallibility. I would hardly regard a person's belief that he is not peculiar as an act of conceit; indeed, to assume that one is not peculiar is a perfectly reasonable assumption. I'm not even sure whether it is psychologically possible for a person to be peculiar. Could a person really believe something and also believe that he doesn't believe it? I doubt it. Yet it is not *logically* impossible for a person to be peculiar.

At any rate, to have confidence in one's own nonpeculiarity is far more reasonable than to have confidence in one's complete accuracy. Therefore the following theorem is a bit sad.

Theorem 3. Suppose a native says to a reasoner of type 1*: "You will never believe that I am a knight." Then if the reasoner believes that he is not (and never will be) peculiar, he will become peculiar!

Problem 5. Prove Theorem 3.

Solution. The proof of this is a bit more elaborate than any other proof so far.

Assume that the native makes this statement and that the reasoner believes that he is incapable of being peculiar. Since the native made the statement, then by Lemma 1, the reasoner will believe that if he ever believes that the native is a knight, he will also believe what the native said. And so the reasoner reasons: "Suppose I should ever believe that he is a knight. Then I'll believe what he says—i.e., I'll believe that I don't believe he is a knight. And so I will then believe he's a knight and I will also believe that I don't believe he is a knight. This means I will be peculiar. Therefore, if I ever believe he's a knight, I will become peculiar. Since I will never be peculiar [sic],

then I will never believe that he's a knight. Since he said I wouldn't, his statement is true, and so he is a knight."

At this point the reasoner has come to the conclusion that the native is a knight, and a bit earlier he came to the conclusion that he doesn't believe that the native is a knight. He has thus lapsed into peculiarity.

Of course the above proposition can be stated and proved in the following more general form:

Theorem B. For any reasoner of type 1*, if he believes any proposition of the form p≡~Bp ("p is true if and only if I will never believe p"), then he cannot believe that he is not peculiar, unless he lapses into peculiarity.

Moral. If you are a reasoner of type 1*, and you wish to believe that you are not peculiar, you can avoid becoming peculiar by simply refusing to believe any proposition of the form "p if and only if I will never believe p."

In particular, if you ever visit the knight-knave island (or what you have been told is a knight-knave island) and a native tells you that you will never believe that he is a knight, then your wisest course is to refuse to believe that the rules of the island hold.

Later in this book, however, when we come to the study of mathematical systems and talk about provability in the system rather than beliefs of a reasoner, we will see that the analogue of the option of not believing p≡~Bp will not be open.

• *Part V* •

THE CONSISTENCY PREDICAMENT

• 11 •

Logicians Who Reason About Themselves

WE ARE getting close to Gödel's consistency predicament. But first, we need to consider reasoners of higher degrees of self-awareness than those of just type 1.

ADVANCING STAGES OF SELF-AWARENESS

We will now define reasoners of types 2, 3, and 4, which represent advancing degrees of self-awareness. Reasoners of type 4 play a major role in the dramas that will unfold.

Reasoners of Type 2. Suppose a reasoner of type 1 believes p and believes p⊃q. Then he will believe q. This means that the proposition (Bp&B(p⊃q))⊃Bq is *true* for a reasoner of type 1. However, the reasoner doesn't necessarily *know* that this proposition is true. Well, we define a reasoner to be of type 2 if he is of type 1 and *believes* all propositions of the form (Bp&B(p⊃q))⊃Bq. (He "knows" that his set of beliefs—past, present, and future—is closed under modus ponens. For any propositions p and q, he believes: "If I should ever

believe p and believe p⊃q, then I will also believe q.")

To emphasize a point, reasoners of type 2 have a certain "self-awareness" not necessarily present in reasoners of type 1. A reasoner of type 1 who believes p and believes p⊃q will sooner or later believe q; reasoners of type 2 also *know* that if they ever believe p and believe p⊃q, they will believe q.

Reasoners of Type 3. We will say that a reasoner is *normal* if for any proposition p, if he believes p, then he believes that he believes p. (If he believes p, then he also believes Bp.) By a reasoner of type 3, we shall mean a normal reasoner of type 2.

Reasoners of type 3 have one more stage of self-awareness than those of type 2.

Reasoners of Type 4. A normal reasoner doesn't necessarily know that he is normal. If a reasoner is normal, then for any proposition p, the proposition Bp⊃BBp is *true* (if he believes p, then he believes Bp), but the reasoner is not necessarily aware of the truth of Bp⊃BBp. Well, we will say that a reasoner *believes* that he is normal if for every proposition p, he believes the proposition Bp⊃BBp. (For every proposition p, the reasoner believes: "If I should ever believe p, then I will believe that I believe p.")

A reasoner of type 3 is in fact normal. By a reasoner of type 4, we mean a reasoner of type 3 who *knows* that he is normal. Thus for any proposition p, a reasoner of type 4 believes the proposition Bp⊃BBp.

As we have remarked, reasoners of type 4 play a major role in this book. Let us review the conditions defining a reasoner of type 4.

(1a) He believes all tautologies.

(1b) If he believes p and believes p⊃q, then he believes q.

(2) He believes (Bp&B(p⊃q))⊃Bq.

(3) If he believes p, then he believes Bp.

(4) He believes Bp⊃BBp.

SOME BASIC PROPERTIES OF SELF-AWARE REASONERS

We will now establish a few basic properties of reasoners of types 2, 3, and 4 that will be used throughout the remaining chapters.

First, a simple observation about reasoners of type 2: For any proposition p and q, the proposition (Bp&B(p⊃q))⊃Bq is logically equivalent to the proposition B(p⊃q)⊃(Bp⊃Bq)—because for *any* propositions X, Y, and Z, the proposition (X&Y)⊃Z is logically equivalent to Y⊃(X⊃Z), as the reader can easily verify—and so any reasoner of type 2 believes all propositions of the form B(p⊃q)⊃(Bp⊃Bq). Conversely, any reasoner of type 1 who believes all propositions of the form B(p⊃q)⊃(Bp⊃Bq) must be of type 2. Let us record this as Fact 1.

Fact 1. A reasoner of type 1 is of type 2 if and only if he believes all propositions of the form B(p⊃q)⊃(Bp⊃Bq).

Suppose now a reasoner of type 2 believes B(p⊃q). He also believes B(p⊃q)⊃(Bp⊃Bq), according to Fact 1, and being of type 1 (since he is of type 2) he will then believe Bp⊃Bq—which is a logical consequence of B(p⊃q) and B(p⊃q)⊃(Bp⊃Bq). And so as an obvious consequence of Fact 1 we have the following corollary: If a reasoner of type 2 believes B(p⊃q), then he will believe Bp⊃Bq.

Now we come to some less obvious facts about reasoners of type 2.

1

Show that for any reasoner of type 2 and any propositions p, q, and r:

(a) He will believe B(p⊃(q⊃r))⊃(Bp⊃(Bq⊃Br)).

(b) If he ever believes B(p⊃(q⊃r)), then he will believe Bp⊃(Bq⊃Br).

2

Show that if a reasoner of type 3 believes $p \supset (q \supset r)$, then he will believe $Bp \supset (Bq \supset Br)$. This fact will have many applications.

3 · Regularity

In the last chapter we defined a reasoner to be of type 1* if he is of type 1 and if for any propositions p and q, if he ever believes $p \supset q$, he will also believe $Bp \supset Bq$. This second condition will be given a name—we will call a reasoner *regular* if his belief in $p \supset q$ implies his belief in $Bp \supset Bq$.

Prove that every reasoner of type 3 is regular (and is thus of type 1*).

4

Prove that if a regular reasoner of type 1 believes $p \equiv q$, then he will believe $Bp \equiv Bq$.

5

There is an interesting connection between regularity and normality. For a regular reasoner of type 1, if there is so much as one proposition q such that he believes Bq, then he must be normal. Why is this?

6

Any reasoner of type 1 who believes p and believes q will believe p&q (which is a logical consequence of the two propositions p and q). Thus the proposition $(Bp \& Bq) \supset B(p \& q)$ is *true* for any reasoner of type 1, hence true for any reasoner of type 3.

Prove that any reasoner of type 3 *believes* (Bp&Bq)⊃B(p&q). (He *knows* that if he should ever believe p and believe q, he will believe p&q.)

CONSISTENCY

We say that a reasoner is *consistent* if the set of all propositions that he believes (or has believed or will believe) is a consistent set, and we shall say that he is *inconsistent* if his set of beliefs is inconsistent. For any reasoner of type 1, the set of his beliefs is logically closed; hence it follows from Principle C of Chapter 8 that the following three conditions are equivalent:

(1) He is inconsistent (he believes ⊥).

(2) He believes some proposition p and its negation (~p).

(3) He believes all propositions.

We shall say that a reasoner *believes* he is consistent if he believes ~B⊥ (he believes that he doesn't believe ⊥). We shall say that he *believes* that he is inconsistent if he believes B⊥ (he believes that he believes ⊥). A reasoner of even type 1 who is inconsistent will also believe that he is inconsistent (because he will believe everything), although a reasoner who believes that he is inconsistent is not necessarily inconsistent (although it can be shown that he must have at least one false belief).

Any reasoner of type 1 who believes some proposition p and its negation ~p will be inconsistent—he will believe ⊥—and so the proposition (Bp&B~p)⊃B⊥ is *true* for a reasoner of type 1.

7

Prove that any reasoner of type 3 *believes* the proposition (Bp&B~p)⊃B⊥.

Note: This last problem is quite crucial for the next chapter. It means that for any proposition p, a reasoner of type 3 *knows* that if he should ever believe p and also believe ~p, then he will be inconsistent. (Of course this also applies to a reasoner of type 4, since every reasoner of type 4 is also of type 3.)

Inconsistency and Peculiarity. We recall that a reasoner is called peculiar if he believes some proposition p and also believes that he doesn't believe p. We have remarked that a peculiar reasoner is not necessarily inconsistent. However, any peculiar reasoner of type 3 *is* inconsistent, as the following problem will reveal.

8

Prove that any peculiar normal reasoner of type 1 must be inconsistent (and hence any peculiar reasoner of type 3 must be inconsistent).

Exercise 1. According to the above problem, for any proposition p, the proposition $(Bp\&B\sim Bp)\supset B\bot$ is *true* for a reasoner of type 3, hence also true for a reasoner of type 4. Prove that any reasoner of type 4 correctly *believes* the proposition $(Bp\&B\sim Bp)\supset B\bot$ (he *knows* that if he should ever be peculiar, he will be inconsistent).

9 · A Little Puzzle

Suppose a reasoner of type 4 believes $p\equiv Bq$. Will he necessarily believe $p\supset Bp$?

AWARENESS OF SELF-AWARENESS

Reasoners of type 4 have one marvelous property not shared by reasoners of lower types—namely, that they *know* they are of type

4, in a sense we will precisely define. Thus, for example, a reasoner may be of type 3 without knowing it, but a reasoner cannot be of type 4 unless he knows it.

Let us say that a reasoner *believes* he is of type 1 if he believes all propositions of the form BX, where X is any tautology, and believes all propositions of the form $(Bp\&B(p\supset q))\supset Bq$. If he also believes all propositions of the form $B((Bp\&B(p\supset q))\supset Bq)$, then we will say that he *believes* he is of type 2. If he also believes all propositions of the form $Bp\supset BBp$, then we will say that he *believes* he is of type 3. If he also believes all propositions of the form $B(Bp\supset BBp)$, then we will say that he *believes* he is of type 4. For each of these types, we will say that a reasoner *knows* that he is of that type if he believes he is of that type and really is of that type.

It is not difficult to see that a reasoner who knows that he is of type 1 is of type 2, and that any reasoner of type 3 knows that he is of type 2 (though he doesn't necessarily know that he is of type 3). Also a reasoner is of type 4 if and only if he knows that he is of type 3. The reader should try proving these facts as exercises.

The following problem is more interesting.

10

Prove that a reasoner of type 4 knows that he is of type 4.

This problem is interesting for several reasons. For one thing, it shows that being of type 4 constitutes a natural resting place in our hierarchy of reasoners. (It would be pointless, for example, to define a reasoner to be of type 5 if he is a reasoner of type 4 who knows that he is of type 4, since any reasoner of type 4 already knows he is of type 4, and thus we wouldn't get anything new.)

Secondly, anything that you or I can prove about all reasoners of type 4, using just propositional logic, any reasoner of type 4 can prove about himself, since he also knows propositional logic and knows that he is of type 4.

Exercise 2. To say that a reasoner is regular is to say that for any propositions p and q, the proposition $B(p \supset q) \supset B(Bp \supset Bq)$ is *true* of the reasoner. (The proposition $B(p \supset q) \supset B(Bp \supset Bq)$ is the proposition that if the reasoner believes $p \supset q$, then he will believe $Bp \supset Bq$.) Let us say that a reasoner *believes* that he is regular if he *believes* all propositions of the form $B(p \supset q) \supset B(Bp \supset Bq)$.

Prove that every reasoner of type 4 knows that he is regular (i.e., he is regular and believes that he is regular).

Exercise 3. Prove that if a reasoner of type 4 believes $p \supset (Bp \supset q)$, then he will believe $Bp \supset Bq$. (The solution to this exercise will be given in Chapter 15; see page 126.)

SOLUTIONS

1 · Suppose the reasoner is of type 2. Take any propositions p, q, and r.

By Fact 1, he believes the following: (1) $B(p \supset (q \supset r)) \supset (Bp \supset B(q \supset r))$. The reason is that for any propositions X and Y, he believes $B(X \supset Y) \supset (BX \supset BY)$, so take p for X and $(p \supset q)$ for Y.

Again, by Fact 1, he believes: (2) $B(q \supset r) \supset (Bq \supset Br)$.

The following proposition is a logical consequence of (2):

(3) $(Bp \supset B(q \supset r)) \supset (Bp \supset (Bq \supset Br))$. The reason is that, for any propositions X, Y, and Z, the proposition $(X \supset Y) \supset (X \supset Z)$ is a logical consequence of $Y \supset Z$, as the reader can verify. We take Bp for X, $B(q \supset r)$ for Y, and $Bq \supset Br$ for Z, and we see that (3) is a logical consequence of (2).

We now know that the reasoner believes both (1) and (2), and $B(p \supset (q \supset r)) \supset (Bp \supset (Bq \supset Br))$ is a logical consequence of (1) and (2), so the reasoner believes it. This proves (a).

(b) Suppose the reasoner believes $B(p \supset (q \supset r))$. By (a) he also believes $B(p \supset (q \supset r)) \supset (Bp \supset (Bq \supset Br))$; hence he will believe $Bp \supset (Bq \supset Br)$, since it is a logical consequence of the two propositions above.

Note: I will not give such detailed arguments in the future. By now the reader should have enough experience to follow briefer arguments and supply missing steps.

2 · Consider now a reasoner of type 3 who believes $p \supset (q \supset r)$. Since he is normal, he will believe $B(p \supset (q \supset r))$. Then, by (b) of the last problem, he will believe $Bp \supset (Bq \supset Br)$.

3 · Suppose a reasoner of type 3 believes $p \supset q$. Since he is normal, he will then believe $B(p \supset q)$. Then, by the corollary to Fact 1, he will believe $Bp \supset Bq$. This proves that he is regular.

4 · Suppose a regular reasoner of type 1 believes $p \equiv q$. Then he will believe both $p \supset q$ and $q \supset p$ (since they are both logical consequences of $p \equiv q$). Being regular, he will then believe both $Bp \supset Bq$ and $Bq \supset Bp$. Then, being of type 1, he will believe $Bp \equiv Bq$ (which is a logical consequence of the last two propositions).

5 · Consider a regular reasoner of type 1. Suppose q is some proposition such that the reasoner believes Bq. Now, let p be any proposition that the reasoner believes. We are to show that he will believe Bp.

The proposition $p \supset (q \supset p)$ is a tautology (as the reader can verify), hence the reasoner believes it. He also believes p (by assumption), hence he will believe $q \supset p$. Then, since he is regular, he will believe $Bq \supset Bp$. Then, since he believes Bq, he will believe Bp. This proves that he is normal.

6 · We are considering a reasoner of type 3. Now, the proposition $p \supset (q \supset (p \& q))$ is obviously a tautology, hence the reasoner will believe it. Then, by (b) of Problem 1, he will believe $Bp \supset (Bq \supset B(p \& q))$, hence he will believe the logically equivalent proposition $(Bp \& Bq) \supset B(p \& q)$. (For any proposition X, Y, and Z, the proposition $X \supset (Y \supset Z)$ is logically equivalent to $(X \& Y) \supset Z$.)

7 · The proposition $(p\&\sim p)\supset\bot$ is a tautology, hence any reasoner of type 3 (or even of type 1) will believe it. Since a reasoner of type 3 is regular (by Problem 3), he will then believe $B(p\&\sim p)\supset B\bot$. He also believes $(Bp\&B\sim p)\supset B(p\&\sim p)$ (by Problem 6). Believing these last two propositions, he will believe $(Bp\&B\sim p)\supset B\bot$, which is a logical consequence of them.

Remarks. For *any* proposition q, the proposition $p\supset(\sim p\supset q)$ is a tautology. Hence by the above argument applied to q, instead of $\bot$, a reasoner of type 3 will believe $(Bp\&B\sim p)\supset Bq$.

8 · This is pretty obvious. If a normal reasoner believes p, he will believe Bp. If he also believes $\sim$Bp and is of type 1, he will be inconsistent. Thus if a normal reasoner of type 1 believes p and believes $\sim$Bp, he will be inconsistent.

Solution to Exercise 1. A reasoner of type 4 believes $Bp\supset BBp$. He thus believes its logical consequence $(Bp\&B\sim Bp)\supset(BBp\&B\sim Bp)$. He also believes $(BBp\&B\sim Bp)\supset B\bot$ (by Problem 7, since $\sim$Bp is the negation of Bp). The proposition $(Bp\&B\sim p)\supset B\bot$ is a logical consequence of the last two propositions, and so the reasoner will believe it.

In other words, a reasoner of type 4 can reason thus: "Suppose I ever believe p and also believe $\sim$Bp. Since I will believe p, I will also believe Bp, hence I will believe both Bp and $\sim$Bp, and then I will be inconsistent. And so, if I ever believe p and $\sim$Bp, I will be inconsistent."

9 · Suppose a reasoner of type 4 believes $p\equiv Bq$. He is regular (by Problem 3, since he is also of type 3), and so by Problem 3 he will believe $Bp\equiv BBq$. Hence he will believe $BBq\supset Bp$. He also believes $Bq\supset BBq$ (since he is of type 4), hence by propositional logic he will believe $Bq\supset Bp$. From $p\equiv Bq$ and $Bq\supset Bp$, he will deduce $p\supset Bp$. And so the answer is yes.

10 · Suppose the reasoner is of type 4. He then satisfies all the conditions that define a reasoner of type 4. We are to show that he *believes* all these conditions.

(1a) Take any tautology X. Being of type 4 (and hence of type 1), he believes X. Then, since he is normal, he believes BX. Thus for any tautology X, he believes BX.

(1b) It follows from the fact that he is of type 2 that he believes all propositions of the form (Bp&B(p⊃q))⊃Bq.

At this point we realize that he knows that he is of type 1. (In fact, by the above argument, any normal reasoner of type 2—i.e., any reasoner of type 3—knows that he is of type 1.)

(2) Since he believes all propositions of form (Bp&B(p⊃q))⊃Bq, and he is normal, then he believes B((Bp&B(p⊃q))⊃Bq).

At this point we see that he knows he is of type 2. (In fact, any reasoner of type 3 knows that he is of type 2.)

(3) Since the reasoner is of type 4, then it is immediate that he knows all propositions of the form Bp⊃BBp—i.e., he knows that he is normal.

At this point we see that a reasoner of type 4 knows that he is of type 3. (But a reasoner of type 3 doesn't necessarily know that he is of type 3, because he may not know that he is normal.)

(4) Since the reasoner of type 4 believes Bp⊃BBp, and he is normal, he believes B(Bp⊃BBp).

Now we see that a reasoner of type 4 knows that he is of type 4 (that is, he knows all of the propositions characterizing a reasoner of type 4).

Solution to Exercise 2. Consider a reasoner of type 4. Since he is of type 1, he believes B(p⊃q)⊃(Bp⊃Bq). Since he is regular, he then believes BB(p⊃q)⊃B(Bp⊃Bq). He also believes B(p⊃q)⊃BB(p⊃q), since he knows he is normal. From this and the last fact, it follows that he must believe B(p⊃q)⊃B(Bp⊃Bq).

• 12 •

The Consistency Predicament

Now the stage is set, and the real show begins!

A reasoner of type 4 visits the Island of Knights and Knaves and believes the rules of the island. And so whenever a native makes a statement, the reasoner will believe that if the native is a knight, the statement is true, and its converse. Thus, if a native asserts a proposition p, then the reasoner will believe $k \supset p$ (where k is the proposition that the native is a knight), and he will also believe $p \supset k$. Moreover, since the reasoner is of type 4, he is regular (as we showed in Problem 3 of the last chapter), and so if the native asserts p, the reasoner will believe not only $k \supset p$, but also $Bk \supset Bp$—that is, he will believe: "If I ever believe that he is a knight, then I will believe what he said."

We recall from the last chapter (Problem 7) that any reasoner of type 4, or even of type 3, knows that if he should ever believe p and believe $\sim p$, he will be inconsistent (p can be any proposition).

Let us review and label these facts.

Fact 1. Suppose a native makes a statement to a reasoner of type 4. Then, (a) The reasoner will believe that if the native is a knight, the statement must be true (and conversely, that if the statement is true, then the native must be a knight). (b) The reasoner will also believe that if he should ever believe that the native is a knight, then he will believe what the native said.

Fact 2. For any proposition p, a reasoner of type 4 knows that if he should ever believe p and also believe ~p, then he will be inconsistent.

With these facts in mind, we are ready to embark. The first big result to which we turn is the following theorem.

***Theorem 1* (after Gödel's Consistency Theorem).** Suppose a native of the island says to a reasoner of type 4: "You will never believe that I am a knight." Then if the reasoner is consistent, he can never know that he is consistent; or, stated otherwise, if the reasoner ever believes that he cannot be inconsistent, he will become inconsistent!

1

Prove Theorem 1.

Solution. Suppose the reasoner does believe that he is (and will remain) consistent. We will show that he will become inconsistent.

The reasoner reasons: "Suppose I ever believe that the native is a knight. Then I'll believe what he said—I'll believe that I don't believe that he is a knight. But also, if I believe he's a knight, then I'll believe that I *do* believe he's a knight (since I am normal). Therefore, if I ever believe that he's a knight, then I'll believe both that I do believe he's a knight and that I don't believe he's a knight, which means I will be inconsistent. Now, I'll never be inconsistent [sic!], hence I will never believe he's a knight. He said that I would never believe he's a knight, and what he said was true, hence he is a knight."

At this point, the reasoner believes the native is a knight, and since he is normal, he will then know that he believes this. Hence the reasoner will continue: "Now I believe he is a knight. He said that I never would, hence he made a false statement, so he is not a knight."

At this point the reasoner believes that the native is a knight and also believes that the native is not a knight, and so he is now inconsistent.

Discussion. The important mathematical content of the above theorem can be presented without reference to knights and knaves. The function of the knight-knave island was to provide a simple method of getting some proposition k (in this case, that the native is a knight) such that the reasoner would believe "k is true if and only if I will never believe k." Any other method of getting such a proposition k would serve as well. Thus Theorem 1 is but a special case of the following theorem (which has nothing to do with knights and knaves).

Theorem G. If a consistent reasoner of type 4 believes some proposition of the form $p \equiv \sim Bp$, then the reasoner can never know that he is consistent. Stated otherwise, if a reasoner of type 4 believes $p \equiv \sim Bp$ and believes that he is (and will remain) consistent, then he will become inconsistent.

We shall prove Theorem G in a sharper form.

Theorem $G^{\#}$. Suppose a normal reasoner of type 1 believes a proposition of the form $p \equiv \sim Bp$. Then:

(a) If he ever believes p, he will become inconsistent.

(b) If he is of type 4, then he knows that if he should ever believe p, then he will become inconsistent—i.e., he will believe the proposition $Bp \supset B\bot$.

(c) If he is of type 4 and believes that he cannot be inconsistent, then he will become inconsistent.

Proof. (a) Suppose he believes p. Being normal, he will then believe Bp. Also, since he believes p and believes $p \equiv \sim Bp$, he must believe $\sim Bp$ (since he is of type 1). And so he will then believe both Bp and $\sim Bp$, hence he will be inconsistent.

(b) Suppose he is of type 4. Since he is of type 1 and believes $p \equiv \sim Bp$, he must also believe $p \supset \sim Bp$. Also, he is regular, hence he will then believe $Bp \supset B \sim Bp$. He also believes $Bp \supset BBp$ (since he knows he is normal). Hence he will believe $Bp \supset (BBp \& B \sim Bp)$, which is a logical consequence of the last two propositions. He also believes $(BBp \& B \sim Bp) \supset B\bot$ (by Fact 2, since for any proposition X, he believes $(BX \& B \sim X) \supset B\bot$, and so he believes this in the special case where X is Bp). Once he believes both $Bp \supset (BBp \& B \sim Bp)$ and $(BBp \& B \sim Bp) \supset B\bot$, he will have to believe $Bp \supset B\bot$ (since he is of type 1).

(c) Since he believes $Bp \supset B\bot$ (as we have just proved), then he also believes $\sim B\bot \supset \sim Bp$. Now, suppose he believes $\sim B\bot$ (he believes he cannot be inconsistent). Since he also believes $\sim B\bot \supset \sim Bp$ (as we have just seen), then he will believe $\sim Bp$. Since he also believes $p \equiv \sim Bp$, he will believe p, hence he will be inconsistent by (a).

The Student and His Theology Professor. Let us now turn again to the student and his theology professor who says to him: "God exists if and only if you will never believe that God exists." If the student believes the professor, then he believes the proposition $g \equiv \sim Bg$, where g is the proposition that God exists. Then, according to Theorem G, the student cannot believe in his own consistency without becoming inconsistent.

We hinted at this at the end of Chapter 2, but we were not then able to state what constituted a "reasonable" set of assumptions about the student's reasoning abilities. Now we can do this. The assumptions are simply that the student is a reasoner of type 4.

Of course the student has the option of believing in his own consistency without becoming inconsistent; he can simply refuse to believe the professor!

Exercise 1. Suppose that in Theorem 1 we are given the additional information that the rules of the island really do hold. Is the native a knight or a knave?

Exercise 2. In the example of the student and his theology professor, suppose that the student does believe the professor and also believes in his own consistency. If God really exists, then was the professor's statement true or false? If God doesn't exist, then was the professor's statement true or false?

Answer to Exercise 1. By Theorem 1, the reasoner will be inconsistent, hence will believe everything. In particular, he will believe that the native is a knight. Since the native said he wouldn't, then the native's statement was false. Therefore, if the rules of the island really hold, the native must be a knave.

Answer to Exercise 2. Again, the student will become inconsistent and believe everything. In particular, he will believe that God exists (he will also believe that God doesn't exist, but this is not relevant here). Letting g be the proposition that God exists, the proposition Bg is thus true, hence ~Bg is false. If God does exist, then g is true, hence g≡~Bg is false, and the professor was wrong. If God doesn't exist, then g is false, g≡~Bg is true, and the professor was right.

Exercise 3. We have proved in earlier chapters the following three facts:

(1) If a native says to a reasoner of type 1*, "You will never believe I'm a knight," and the reasoner believes he will never be peculiar, then he will become peculiar. (Theorem 3, Chapter 10, page 84.)

(2) A peculiar reasoner of type 3 is inconsistent. (Problem 8, Chapter 11, page 94.)

(3) A reasoner of type 4 believes that if he should become peculiar, he will be inconsistent. (Exercise 1, Chapter 11, page 94.)

Using these three facts, one can give a much swifter proof of Theorem 1 of this chapter (page 101) than the one we have given. How?

Answer to Exercise 3. Suppose the native makes this statement—"You will never believe that I am a knight"—to a reasoner of type

4 and the reasoner believes that he will never be inconsistent. Then the reasoner will believe that he cannot be peculiar (because he knows that if he is peculiar, he will be inconsistent). Then by Fact 1 above, he will become peculiar. And by Fact 2 above, he will become inconsistent.

THE DUAL OF THEOREM G

Exercise 4. Suppose that the native, instead of saying, "You will never believe I'm a knight," says, "You *will* believe I'm a knave." Assuming the reasoner is of type 4 and believes in his own consistency, does it now follow that he will become inconsistent?

Solution. The answer is yes, and this can be easily established as a corollary of Theorem G, but first I'd like to sketch a direct argument. The reasoner reasons: "Suppose he's a knight. Then what he said is true, which means that I'll believe he is a knave. Once I believe he's a knave, I'll believe the *opposite* of what he said—I'll believe that I *don't* believe he's a knave. But if I believe he's a knave, I'll also believe that I *do* believe he's a knave (because I'm normal), and hence I'll be inconsistent. Since I will never be inconsistent, he can't be a knight after all; he must be a knave. Now I believe he's a knave. He said I would, and so he's a knight."

At this point the reasoner is inconsistent.

What we have just proved is a special case of the following "dual" of Theorem G.

Theorem G°. If a consistent reasoner of type 4 believes some proposition of the form $p \equiv B \sim p$, then he can never know that he is consistent.

2

Theorem G° can be proved by "dualizing" the argument for Theorem G, but it can be obtained much more simply as a corollary of Theorem G. How?

Solution. Suppose the reasoner believes $p \equiv B\sim p$. Then he will believe $\sim p \equiv \sim B\sim p$. Let q be the proposition $\sim p$. Then the reasoner believes $q \equiv \sim Bq$, and so the reasoner does believe a proposition of the form $p \equiv \sim Bp$ (namely, $q \equiv \sim Bq$), and so the result follows by Theorem G.

Exercise 5. Suppose that in Exercise 4 we add the assumption that the rules of the island really hold and that the reasoner *does* believe in his own consistency. Is the native a knight or a knave?

Exercise 6. Suppose a theology professor says to his student of type 4: "God exists if and only if you will believe that God doesn't exist." Suppose the student believes the professor and also believes in his own consistency. Prove that if the professor's statement is true, then God must exist. Prove that if the professor's statement is false, then God doesn't exist.

Exercise 7. Suppose a native tells a reasoner of type 4: "You will never believe I'm a knight" (or, alternatively, says: "You will believe I'm a knave"). Prove that the reasoner *knows* that if he should ever believe in his own consistency, he will become inconsistent. (The solution of this will follow easily from results to be proved in later chapters.)

• 13 •

Gödelian Systems

ALL THE results we have so far proved about reasoners and their beliefs are counterparts of metamathematical results about mathematical systems and the propositions provable in them. Before turning to these, let us summarize the most significant facts we have proved in the last few chapters.

Summary I. Suppose a reasoner believes the rules of the island and a native says to him: "You will never believe that I am a knight." Or, more generally, suppose a reasoner believes some proposition of the form $p \equiv \sim Bp$ (p is true if and only if I will never believe p). Then:

(1) For a reasoner of type 1, if he believes that he is always accurate, then he will become inaccurate; in fact, he will become peculiar.

(2) For a reasoner of type 1*, if he believes that he will never be peculiar, then he will become peculiar.

(3) For a reasoner of type 3, if he believes that he will never be peculiar, then he will become inconsistent.

(4) For a reasoner of type 4, if he believes that he will always be consistent, then he will become inconsistent.

All these facts, particularly (4), are related to important metamathematical facts, which we will discuss briefly.

The types of systems investigated by Kurt Gödel have the following features. First, there is a well-defined set of propositions expressible in the system; these will be called the propositions *of* the system. One of these propositions is $\perp$ (logical falsehood), and for any proposition p and q of the system, the proposition $(p \supset q)$ is also a proposition of the system. The logical connectives &, v, $\sim$, $\equiv$ can all be defined from $\supset$ and $\perp$ in the manner explained in Chapter 7.

Second, the system—call it "S"—has various axioms and logical rules making certain propositions *provable* in the system. We thus have a well-defined subset of the set of propositions of the system—namely, the set of *provable* propositions of the system.

Third, for any proposition p of the system, the proposition that p is *provable* in the system is itself a proposition *of* the system (it may be true or false, and it may be provable in the system, or then again it may not). We let Bp be the proposition that p is provable in the system. (The symbol "B" for "provable" was introduced by Gödel; it stands for the German word *beweisbar.*) By a fortunate coincidence, we have the symbol "B" used in two closely related situations. When we speak of *reasoners,* Bp means that the reasoner believes p; when we speak of mathematical *systems,* Bp means that p is *provable* in the system.

We now define a system S to be of type 1, 1*, 2, 3, 4, in exactly the same way as we did for reasoners: If all tautologies are provable in S, and if for any propositions p and $p \supset q$ both provable in S, q is also provable in S—if these two conditions hold—then we say that S is of type 1. If also $Bp \supset Bq$ is provable in S whenever $p \supset q$ is provable in S, then we say that S is of type 1*. If also all propositions of the form $(Bp \& B(p \supset q)) \supset Bq$ are provable in S, then we say that S is of type 2. If also S is normal (i.e., Bp is provable whenever p is), then we say that S is of type 3. Lastly, if all propositions of the form $Bp \supset BBp$ are provable in S, then we say that S is of type 4. Of course, all the results of Chapter 11 that we proved for *reasoners* hold good

also for *systems* (where we now reinterpret "B" to mean *provable* rather than *believed*). As for the results of Chapters 10 and 12, we need another condition to which we now turn.

Gödelian Systems. Gödel made the remarkable discovery that each of the systems which he investigated had the property that there was a proposition p such that the proposition $p \equiv \sim Bp$ was provable in the system. Such systems we will call *Gödelian* systems.

The proposition $p \equiv \sim Bp$ is a very curious one. Here we have a proposition p equivalent to its own nonprovability in the system! The proposition p can be thought of as saying, "I am not provable in the system." How Gödel managed to find such a proposition need not concern us now, although it will be taken up in a much later chapter.

In analogy to systems, we might define a reasoner to be a Gödelian reasoner if there is at least one proposition p such that he believes the proposition $p \equiv \sim Bp$. Of course we have been studying Gödelian reasoners throughout the last chapter. (If a reasoner comes to the Island of Knights and Knaves and believes the rules of the island, and if a native says to him, "You will never believe that I am a knight," then the reasoner believes the proposition $k \equiv \sim Bk$, hence he becomes a Gödelian reasoner.)

Let us now say that a system can prove its own accuracy if it can prove all propositions of the form $Bp \supset p$. We will say that it can prove its own nonpeculiarity if it can prove all propositions of the form $\sim(Bp \& B\sim Bp)$. We will say that the system can prove its own consistency if it can prove the proposition $\sim B\bot$.

Let us now restate our opening summary in terms of systems, rather than reasoners.

Summary I*. (1) If a Gödelian system of type 1 can prove its own accuracy, then it is inaccurate—in fact, peculiar.

(2) If a Gödelian system of type 1* can prove its own nonpeculiarity, then it is peculiar.

(3) If a Gödelian system of type 3 can prove its own non-peculiarity, then it is inconsistent.

(4) If a Gödelian system of type 4 can prove its own consistency, then it is inconsistent.

Item (4) above is the really important one; it is a generalized form of Gödel's famous Second Incompleteness Theorem.

Discussion. In Gödel's original 1931 paper, he took a particular system (the system of *Principia Mathematica* of Whitehead and Russell) and showed that the system, if consistent, couldn't prove its own consistency. He stated that his method applied not only to this particular system, but to a wide variety of systems. Indeed, the method applies to all Gödelian systems of type 4, as we have just seen.

Another important Gödelian system of type 4 is the system known as "Arithmetic" (more completely, "First-Order Peano Arithmetic"). This is formalization of the theory of the ordinary whole numbers 0, 1, 2, . . . Since Arithmetic is a Gödelian system of type 4, it is subject to Gödel's Consistency Theorem; hence if Arithmetic is consistent (which it is, since only true propositions are provable in it), then it cannot prove its own consistency.

When the mathematician André Weil heard about this, he made the famous quip, "God exists, since Arithmetic is consistent; the Devil exists, since we cannot prove it."

This quip, though delightful, is actually misleading. It's not that we can't prove the consistency of Arithmetic; it is that Arithmetic can't prove the consistency of Arithmetic! *We* certainly can prove the consistency of Arithmetic, but our proof cannot be formalized in Arithmetic itself.

Indeed, there has been a good deal of popular misunderstanding concerning Gödel's Second Theorem (the Consistency Theorem); this has been partly due to the irresponsibility of some science reporters and other popularizers. (I, of course, am certainly all for popularization, providing the popularization is not inaccurate.) One

particular popularizer wrote: "Gödel's theorem means that we can never know that Arithmetic is consistent." This is sheer nonsense. To see how silly it is, suppose it had turned out that Arithmetic could prove its own consistency—or, to be more realistic, suppose we take some other system that *can* prove its own consistency. Would this be any guarantee of the consistency of the system? Of course not. If the system were inconsistent, then, being of type 1, it could prove *anything,* including its own consistency! To trust the consistency of a system on the grounds that it can prove its own consistency would be as foolish as to trust the veracity of a person on the grounds that he claims to be always truthful. No, the fact that Arithmetic can't prove its own consistency doesn't cast the faintest ray of doubt on the consistency of Arithmetic.

As a matter of fact, in this book we will construct several Gödelian systems of type 4, and we will *prove* their consistency beyond a shadow of a doubt. Then, we will show that by virtue of their very consistency, the systems will be unable to prove their own consistency.

Exercise. Consider a system S of type 4 (but not necessarily Gödelian). Recall that for any proposition p, the proposition Bp is true if and only if p is provable in S.

(a) First show that for any proposition p, the proposition $(B(p \equiv \sim Bp) \& B \sim B\bot) \supset B\bot$ is true (for the system S). This is quite easy.

(b) Then show that $(B(p \equiv \sim Bp) \& B \sim B\bot) \supset B\bot$ is actually provable in S. (Not so easy!)

• 14 •

More Consistency Predicaments

SOME PRELIMINARY PROBLEMS

1

Suppose a reasoner believes that he is inconsistent. Is he necessarily inconsistent? Is he necessarily inaccurate?

2

Suppose a reasoner believes that he is inaccurate. Prove that he is right!

3

Suppose a reasoner of type 1* believes that he cannot be inconsistent (he believes $\sim B\bot$). Will he necessarily believe that he will never believe that he is inconsistent? (I.e., will he necessarily believe $\sim BB\bot$?)

SOME MORE CONSISTENCY PREDICAMENTS

We are back to the Island of Knights and Knaves. We continue to assume that the reasoner is of type 4 and that he believes the rules of the island.

4

Suppose a native says to the reasoner, "If I am a knight, then you will believe that I'm a knave." Prove:

(a) The reasoner will sooner or later believe himself inconsistent.

(b) If the rules of the island really hold, then the reasoner will become inconsistent!

Note: In the problems of this chapter, we are *not* assuming that the reasoner believes that he is consistent.

5

Suppose the native says instead, "If I am a knight, then you will never believe that I am one." Prove:

(a) The reasoner will become inconsistent.

(b) The rules of the island don't really hold.

Note 1: This problem holds good even if the reasoner is only of type 3.

Note 2: For a "student and his theology professor" version of the last two problems, see the discussion following the solution to Problem 5.

6

Here is a curious one. A reasoner of type 4 comes to what he *believes* is a knight-knave island (he believes the rules of the island), and a native says to him the following two things:

(1) "You will believe that I am a knave."

(2) "You will always be consistent."

Prove that the reasoner will become inconsistent and that the rules of the island don't hold.

7

This time a native makes the following two statements to a reasoner of type 4:

(1) "You will never believe I'm a knight."

(2) "If you ever believe I'm a knight, then you will become inconsistent."

Prove that the reasoner will become inconsistent and that the rules of the island don't hold.

TIMID REASONERS

Let us say that a reasoner shouldn't believe p if his believing p will lead him into an inconsistency. Let us say that a reasoner is *afraid* to believe p if he believes $Bp \supset B\bot$—i.e., if he believes that his believing p will lead him into an inconsistency. (In other words, he is afraid to believe p if he believes that he shouldn't believe p.)

We know that any reasoner of type 4 who believes the rules of the island and is told by a native that he will never believe that he is a knight *shouldn't* believe in his own consistency. In general, however, there is no reason why a reasoner of type 4 shouldn't believe in his own consistency. But now arises a very curious thing: If for some reason, a reasoner of type 4 *is* afraid of believing in his own consistency, his very fear justifies it. By this I mean that any

reasoner of type 4 who is afraid of believing in his own consistency, really *shouldn't* believe in his own consistency. Put still another way, if a reasoner of type 4 believes that his believing in his own consistency will get him into an inconsistency, then it will!

Surprising as this fact may be, it is not difficult to prove. Moreover, this fact holds even for normal reasoners of type 1.

8

Prove that if a normal reasoner of type 1 is afraid to believe in his own consistency, then he really shouldn't believe in his own consistency.

Remarks. In the above problem, we see that for a normal reasoner of type 1, his belief in the proposition $B{\sim}B\bot{\supset}B\bot$ is *self-fulfilling* in the sense that his believing that proposition is a sufficient condition for the proposition being true. The theme of self-fulfilling beliefs will play a major role in the next few chapters.

9

A reasoner of type 4 is also a normal reasoner of type 1, hence according to the last problem, if he is afraid to believe in his own consistency, then he shouldn't believe in his own consistency. This means that for a reasoner of type 4, the proposition $B(B{\sim}B\bot{\supset}B\bot){\supset}(B{\sim}B\bot{\supset}B\bot)$ is true. Prove that any reasoner of type 4 *knows* that this proposition is true (he knows that if he is afraid to believe in his own consistency, then he shouldn't believe in it).

10

Suppose a reasoner of type 4 *believes* that he is afraid to believe in his own consistency. Does it follow that he really *is* afraid to believe in his own consistency?

SOLUTIONS

1 · Suppose a reasoner believes that he is inconsistent. I see no reason why he is necessarily inconsistent, but he must be inaccurate for the following reasons.

A reasoner who believes he is inconsistent is either right or wrong in this belief. If he is wrong in this belief, then he is obviously inaccurate (he has the false belief that he is inconsistent). If he is right in this belief, then he really does believe the false proposition $\perp$. In either case, he has at least one false belief.

2 · If he were wrong, then he would be accurate, which is a contradiction.

3 · Suppose he believes $\sim B\perp$. Then he believes the logically equivalent proposition $B\perp \supset \perp$. Also he is regular (since he is of type 1*), and hence he will believe $BB\perp \supset B\perp$, hence he will believe the logically equivalent proposition $\sim B\perp \supset \sim BB\perp$. Since he also believes $\sim B\perp$, then he will believe $\sim BB\perp$.

4 · We know by Theorem 1 of Chapter 3 that for any proposition p, if a native of a knight-knave island says, "If I am a knight then p," the native must be a knight and p must be true. Now, any reasoner—even if type 1—knows this as well as you and I, and so if the reasoner *believes* that the rules of the island hold, then if a native says to him, "If I am a knight, then p," the reasoner will believe that the native is a knight and that p is true. In this particular problem, the native has said, "If I am a knight, then you will believe that I am a knave," and so the reasoner will believe that the native is a knight and also that he (the reasoner) will *believe* that the native is a knave. And so the reasoner will believe k and also believe $B\sim k$ (k is the proposition that the native is a knight). So far, we have used only the fact that the reasoner is of type 1. However, he is of type 4, and since he believes k, he will also believe Bk. Hence he will

believe Bk and believe B∼k, but he knows that (Bk&B∼k)⊃B⊥ (as we showed in Chapter 11, page 98). Therefore he will believe B⊥—i.e., he will believe that he is (or will be) inconsistent. He also believes that the native is a knight.

So far, we have not used the fact that the rules of the island really hold; our preceding argument used only the fact that the reasoner *believes* that the rules hold. Now, suppose the rules really do hold. Then the native really is a knight and the reasoner really will believe that the native is a knave (according to Theorem 1, Chapter 3). But since the reasoner also believes that the native is a knight, he will become inconsistent.

5 · This time the reasoner believes k≡(k⊃∼Bk), hence he believes k and believes ∼Bk, which are logical consequences of k≡(k⊃∼Bk). Since he believes k, he will believe Bk, and believing ∼Bk, he will become inconsistent.

If the rules of the island really hold, then k≡(k⊃∼Bk) is not only believed by the reasoner, but is actually true, hence k&∼Bk (which is logically implied by it) is true, hence ∼Bk is true, contrary to the fact that the reasoner *does* believe the native is a knight (and hence Bk, rather than ∼Bk, is true). Thus the rules of the island don't really hold.

Discussion. Let us look at the last two problems in the form of the student and his theology professor. Suppose the professor says: "God exists, but you will never believe that God exists." If the student is of type 4 and believes the professor, he will have to believe himself inconsistent. If also the professor's statement was true, then the student really will become inconsistent.

On the other hand, suppose the professor says: "God exists, but you will never believe that God exists." Then if the student is of type 4—or even of type 3—and believes the professor, he will become inconsistent, and also, the professor's statement is false.

6 · The reasoner reasons: "Suppose he is a knave. Then his second statement is false, which means that I will be inconsistent, hence I'll believe everything—in particular, that he's a knave. But this will validate his first statement and make him a knight. Therefore it is contradictory to assume that he is a knave, hence he must be a knight. Since he is a knight, his first statement is true, hence I will believe he is a knave. But I now believe he is a knight, hence I will be inconsistent. This proves that I will be inconsistent. However, he's a knight and said that I will always be consistent. Therefore I won't ever be inconsistent."

At this point the reasoner has come to the conclusion that he will become inconsistent and that he won't become inconsistent, and so he is now inconsistent.

Since the reasoner has become inconsistent, the native's second statement is false. Also, since the reasoner has become inconsistent, he will believe everything, including the fact that the native is a knave. This makes the native's first statement true. Since the native has made one true statement and one false statement, then the rules of the island don't really hold.

7 · The reasoner believes the following two statements:

(1) $k \equiv \sim Bk$
(2) $k \equiv (Bk \supset B\bot)$

Since he believes (1), then, according to (b) of Theorem $G^{\#}$, Chapter 12 (page 102), the reasoner will believe $Bk \supset B\bot$. And, since he believes (2), he will believe k. So, according to (a) of Theorem $G^{\#}$, Chapter 12, he will become inconsistent.

It further follows that the native's first statement was false and his second statement true. Therefore the rules of the island don't hold.

8 · We assume that the reasoner is normal and of type 1 and that he believes $B{\sim}B\bot \supset B\bot$. We are to show that if he believes he is consistent, he will become inconsistent (and therefore that his fear

is justified). So, suppose he ever does believe $\sim B\bot$. Being normal, he will then believe $B\sim B\bot$. This, together with his belief in $B\sim B\bot \supset B\bot$, will cause him to believe $B\bot$ (because he is of type 1). And so he will believe both $B\bot$ and $\sim B\bot$ (he will believe that he is inconsistent and that he is not inconsistent), which will make him inconsistent.

9 · The reasoner reasons: "Suppose I become afraid to believe in my own consistency. This means that I'll believe that I shouldn't believe in my own consistency—i.e., I'll believe $B\sim B\bot \supset B\bot$. Since I am of type 1 (being of type 4), then I will believe $\sim B\bot \supset \sim B\sim B\bot$. Now suppose I should also believe that I am consistent—i.e., suppose I believe $\sim B\bot$. Then, since I will believe $\sim B\bot \supset \sim B\sim B\bot$, I'll believe $\sim B\sim B\bot$. But I'll also believe $B\sim B\bot$ (since I'll believe $\sim B\bot$ and I am normal), and hence I'll be inconsistent. Therefore, if I am afraid to believe in my own consistency, I really cannot believe in my own consistency without becoming inconsistent. Thus the proposition $B(B\sim B\bot \supset B\bot) \supset (B\sim B\bot \supset B\bot)$ is true."

10 · We have just seen that the reasoner *believes* the proposition $B(B\sim B\bot \supset B\bot) \supset (B\sim B\bot \supset B\bot)$. Therefore, if he believes $B(B\sim B\bot \supset B\bot)$, he will believe $B\sim B\bot \supset B\bot$—which means that if he *believes* that he is afraid to believe in his own consistency, then he really will be afraid to believe in his own consistency.

• *Part VI* •

SELF-FULFILLING BELIEFS AND LÖB'S THEOREM

• 15 •

Self-Fulfilling Beliefs

THE PROBLEMS of this chapter are all related to Löb's Theorem, a famous result that is important to the thrust of this book.

We now have a change of scenario: A reasoner of type 4 is thinking of visiting the Island of Knights and Knaves because he has heard a rumor that the sulfur baths and mineral waters there might cure his rheumatism. Before embarking, however, he discusses the situation with his family physician. He asks the doctor whether the "cure" really works. The doctor replies: "The cure is largely psychological; the belief that it works is self-fulfilling. *If you believe that the cure will work, then it will work.*"

The reasoner trusts his doctor implicitly, and so he goes to the island with the prior belief that if he believes the cure will work, then the cure will work. He takes the cure, which lasts only a day but which is not supposed to work for several weeks, if it works at all. The next day he starts worrying about the situation and thinks: "If only I can *believe* that the cure works, then it will work. But how do I know whether I will ever believe that it works? I have no rational evidence that the cure will work, nor do I have any evidence that I will ever believe that the cure will work. For all I know, I may never believe that the cure will work, and the cure might accordingly not work!"

A native of the island passes by and asks the reasoner why he looks so disconsolate. The reasoner explains the entire situation, then summarizes it by saying: "If I ever believe the cure will work, then it will, but will I ever believe the cure will work?" The native replies: *"If you ever believe I'm a knight, then you will believe that the cure will work."*

At first, this did not seem particularly reassuring to the reasoner. He thinks: "What good does this do me? Even if what he says is true, this will only reduce the problem to whether I will ever believe that he is a knight. How do I know whether I will ever believe he is a knight? And even if I do, he may be a knave and his statement may be false, hence I may still not believe that the cure will work." But then the reasoner thought some more about his problem, and after a while he heaved a sigh of relief. Why?

Well, as we will see, the amazing thing is that the reasoner *will* believe that the cure will work and, assuming that the doctor is right, the cure will work! I might remark that the rules of the island don't have to really hold for the argument to go through; it is enough that the reasoner *believes* that they hold.

This problem is closely related to M. H. Löb's important theorem, which we will examine later. But first, let's consider a slightly simpler problem, one that comes even closer to Löb's original argument. Suppose that the native, instead of saying the above, says: "If you ever believe that I'm a knight, then the cure will work."

1 · (After Löb)

Prove that under the above conditions, the reasoner *will* believe that the cure will work (and hence, if the doctor was right, then the cure will work).

Solution. It will be easiest to give the solution partly in words and partly in symbols. We let k be the proposition that the native is a

knight, and we let C be the proposition that the cure will work. At the outset, the reasoner believes the proposition BC⊃C.

The reasoner reasons: "Suppose I ever believe that he is a knight. Then I'll believe what he says—I'll believe that Bk⊃C. Also, if I ever believe he's a knight, I'll believe that I believe he's a knight—I'll believe Bk. And so, if I ever believe he is a knight, I'll believe both Bk and Bk⊃C, hence I'll believe C. Thus, if I ever believe he is a knight, then I'll believe that the cure works. But if I ever believe that the cure works, then the cure will work (as my doctor told me). And so, if I ever believe he's a knight, then the cure will work. Well, that's exactly what he said. He said that if I ever believe he's a knight, then the cure will work, and he was right! Hence he is a knight."

At this point the reasoner believes that the native is a knight, and since the reasoner is normal, he continues: "Now I believe he is a knight. I have already proved that if I believe he is a knight, then the cure will work, and since I do believe that he is a knight, the cure will work."

At this point the reasoner believes that the cure will work. Then, assuming his doctor was right, the cure will work.

The solution to Problem 1 could have been established more quickly had we first proved the following lemma, which will have other applications as well.

Lemma 1. Given any proposition p, suppose a native of the island says to reasoner of type 4: "If you ever believe that I'm a knight, then p is true." Then the reasoner will believe: "If I ever believe he is a knight, then I will *believe* p." More generally, for any two propositions k and p, if a reasoner of type 4 believes the proposition k≡(Bk⊃p), or even believes the weaker proposition k⊃(Bk⊃p), then he will believe Bk⊃Bp.

Exercise 1. How is Lemma 1 proved? (This is the same problem found in Exercise 3, Chapter 11, page 96.)

Exercise 2. How does the use of Lemma 1 facilitate the solution of Problem 1?

Answer to Exercise 1. Let's first show this in the knight-knave version. We let k be the proposition that the native is a knight. The native has asserted the proposition Bk⊃p. The reasoner reasons: "If I ever believe that he's a knight, then I'll believe what he says—I'll believe Bk⊃p. But if I believe he's a knight, I'll also believe Bk (I'll believe that I believe he's a knight). Once I believe Bk⊃p and I believe Bk, then I'll believe p. And so, if I ever believe he's a knight, then I'll believe p."

Of course the more general form can be proved in essentially the same manner, or alternatively as follows: Suppose a reasoner of type 4 believes k⊃(Bk⊃p), which he will certainly believe, if he believes the stronger proposition k≡(Bk⊃p). Then, according to Problem 2, Chapter 11, he will believe Bk⊃(BBk⊃Bp). He also believes Bk⊃BBk. Believing these two propositions, he will believe Bk⊃Bp, which is a logical consequence of them. (For any proposition X, Y, and Z, the proposition X⊃Z is a logical consequence of X⊃(Y⊃Z) and X⊃Y. In particular, this is so if X is the proposition Bk, Y is the proposition BBk, and Z is the proposition Bp.)

Answer to Exercise 2. The reasoner believes k≡(Bk⊃C)—because the native asserted Bk⊃C. Then, according to Lemma 1, the reasoner will believe Bk⊃BC. He also believes BC⊃C, hence he will believe Bk⊃C. Then he will believe k (since he believes both Bk⊃C and k≡(Bk⊃C)), hence he will believe Bk (he is normal). Now that he believes Bk and believes Bk⊃C, he will believe C.

The upshot of Problem 1 is the following theorem.

Theorem 1 (After Löb). For any proposition k and C, if a reasoner of type 4 believes BC⊃C and believes k≡(Bk⊃C), then he will believe C.

Theorem 1 yields the following curious result.

Exercise 3. Suppose a theology student is worried about such things as the existence of God and his own salvation. He asks his professor: "Does God exist?" And "Will I be saved?" The professor then makes the following statements:

(1) "If you believe that you will be saved, then you will be saved."

(2) "If God exists and you believe that God exists, then you will be saved."

(3) "If God doesn't exist, then you will believe that God exists."

(4) "You will be saved only if God exists."

Assuming that the student is a reasoner of type 4 and that he believes his professor, prove: (a) The student will believe that he will be saved; (b) If the professor's statements are true, then the student will be saved.

Solution. Let g be the proposition that God exists and let S be the proposition that the student will be saved. The student then believes the following four propositions:

(1) $BS \supset S$
(2) $(g \& Bg) \supset S$
(3) $\sim g \supset Bg$
(4) $S \supset g$

The proposition $\sim Bg \supset g$ is a logical consequence of (3). This proposition, together with (4), has as a logical consequence the proposition $(\sim Bg \vee S) \supset g$. Also $\sim Bg \vee S$ is logically equivalent to $Bg \supset S$, hence $(\sim Bg \vee S) \supset g$ is logically equivalent to $(Bg \supset S) \supset g$. Also $g \supset (Bg \supset S)$ is logically equivalent to (2), and $g \equiv (Bg \supset S)$ is a logical consequence of $(Bg \supset S) \supset g$ and $g \supset (Bg \supset S)$. Therefore $g \equiv (Bg \supset S)$ is a logical consequence of (1), (2), and (3). Since the student believes (1), (2), and (3), he will also believe $g \equiv (Bg \supset S)$. Since by (1) he also believes $BS \supset S$, then by Theorem 1, he will believe S. Thus BS is true, and if the professor was right, $BS \supset S$ is true, hence S is true, which means that the student will be saved.

. . .

Now let us return to Problem 1 and its proposition: "If you ever believe I'm a knight, then the cure will work."

2

Suppose a reasoner of type 4 again believes BC⊃C, but this time the native says: "If you ever believe I'm a knight, then you will *believe* that the cure works." Prove that the reasoner will again believe that the cure will work.

Solution. This time the native is asserting Bk⊃BC (instead of Bk⊃C). Then by Lemma 1, the reasoner will believe Bk⊃BBC (instead of Bk⊃BC). However, the reasoner believes BC⊃C, and since he is of type 4, he is regular, hence he will believe BBC⊃BC. Believing this and Bk⊃BBC, he will believe Bk⊃BC. He also believes k≡(Bk⊃BC), so he will believe k. Then he will believe Bk, and since he will believe Bk⊃BC, he will believe BC. But he also believes BC⊃C, hence he will believe C.

Of course the above argument generalizes as follows:

Theorem 2. Given any propositions k and C, if a reasoner of type 4 believes BC⊃C and believes k≡(Bk⊃BC), then he will believe C.

3

Again a reasoner of type 4 has the prior belief that if he believes that the cure will work, then the cure will work. This time the native says to him: "Sooner or later you will believe that if I am a knight, then the cure will work." We will see that again the reasoner will believe that the cure will work.

More generally, we will prove the following theorem.

Theorem 3. For any propositions k and C, if a reasoner of type 4 believes BC⊃C and believes k≡B(k⊃C), then he will believe C.

The proof of Theorem 3 is facilitated by the following two lemmas, which are of interest in their own right.

Lemma 2. Suppose that for some proposition q, a native says to a reasoner of type 4: "You will believe q." Then the reasoner will believe: "If he is a knight, then I will believe that if he is a knight, I will believe that he is a knight." (Stated more abstractly, for any propositions k and q, if a reasoner of type 4 believes $k \equiv Bq$, then he will believe $k \supset Bk$.)

Lemma 3. Given any proposition p, suppose a native says to a reasoner of type 4: "You will believe that if I am a knight, then p is true." Then the reasoner will believe: "If he is a knight, then I will *believe* p." (Stated more abstractly, if a reasoner of type 4 believes $k \equiv B(k \supset p)$, then he will believe $k \supset Bp$.)

How are Lemmas 2 and 3 and Theorem 3 proved?

Proof of Lemma 2. This is Problem 9 of Chapter 11, which we have already solved, but I wish to give a knight-knave version of the proof, which is particularly intuitive. The native has said: "You will believe q." The reasoner then reasons: "Suppose he is a knight. Then I *will* believe q. Then I'll believe that I believe q, hence I'll believe what he said, hence I'll believe he's a knight. Therefore, if he is a knight, then I'll believe he's a knight."

Proof of Lemma 3. We shall use Lemma 2 to facilitate this proof.

The native has said: "You will believe that if I'm a knight, then p." Let q be the proposition "If I'm a knight, then p." The native has told the reasoner that he will believe q, and so by Lemma 2, the reasoner will believe that if the native is a knight, then he (the reasoner) will believe that the native is a knight. And so the reasoner reasons: "Suppose he is a knight. Then I'll believe that he is a knight. Then I'll believe what he says—I'll believe $Bk \supset p$. I'll also believe Bk (I'll believe that I believe he is a knight). Therefore, if he is a knight,

then I'll believe Bk⊃p and I'll believe Bk, hence I will also believe p. And so if he is a knight, then I will believe p."

Proof of Theorem 3. I will give a knight-knave version of the proof. The native has said: "You will believe that if I am a knight, then the cure will work." By Lemma 3, the reasoner will believe: "If he is a knight, then I will believe that the cure will work." The reasoner then continues: "Also, if I believe that the cure will work, then the cure will work. Therefore, if he is a knight, the cure will work. I now believe that if he is a knight, then the cure will work. He said I would believe that, hence he is a knight. And so he is a knight, and in addition (as I have proved), if he is a knight, then the cure will work. Therefore the cure will work."

At this point the reasoner will believe that the cure will work.

Discussion. One can also obtain Theorem 3 as an easy corollary of Theorem 1 by the following argument. Suppose we are given propositions k and C such that a reasoner of type 4 believes k≡B(k⊃C) and believes BC⊃C. We are to show that he will believe C. Since he believes k≡B(k⊃C), he must also believe (k⊃C)≡(B(k⊃C)⊃C)), which is a logical consequence of k≡B(k⊃C)). Then he believes the proposition k′≡(Bk′⊃C), where k′ is the proposition k⊃C. Since he also believes BC⊃C, then he will believe C by Theorem 1.

The following curious exercise illustrates self-reference carried to its extreme.

Exercise 4. Suppose a native of the island says to a reasoner of type 4: "Sooner or later you will believe that if I am a knight, then you will believe that I am one."

(a) Prove that the reasoner will believe that the native is a knight.

(b) Prove that if the rules of the island really hold, then the native is a knight.

Solution. This is an easy consequence of Lemma 2.

(a) The native has asserted B(k⊃Bk). Thus there is a proposition

q—namely, k⊃Bk—such that the native has claimed that the reasoner will believe q. Then by Lemma 2, the reasoner will believe k⊃Bk. Then, since he is normal, he will believe B(k⊃Bk)—he will believe what the native said. Hence he will believe that the native is a knight.

(b) Since the reasoner believes Bk, he certainly believes k⊃Bk, hence the native's statement was true. And so the native is a knight (if the rules of the island really hold).

The essential mathematical content of the above exercise is that for any proposition k, if a reasoner of type 4 believes k≡B(k⊃Bk), then he will also believe k. If also k≡B(k⊃Bk) is true, so is k.

DUAL FORMS

Problems 1, 2, and 3 (more generally, Theorems 1, 2, and 3) have their "dual" forms, which are rather curious.

1° · (Dual of Problem 1)

Again, a reasoner of type 4 comes to the island already believing that if he believes that the cure will work, then the cure will work. He meets a native who says to him: "The cure will not work and you will believe that I am a knave."

Prove that the reasoner will believe that the cure will work.

2° · (Dual of Problem 2)

Like Problem 1°, except that the native says: "You will believe that I'm a knave, but you will never believe that the cure will work." Show that the same conclusion follows (the reasoner will believe that the cure will work).

3° · (Dual of Problem 3)

This time the native says: "You will never believe that if I am a knave, then the cure will work." (Alternatively, he could say: "You will never believe that either I am a knight or that the cure will work.") Show that the same conclusion follows.

Solution to Problem 1°. We could prove this from scratch, but it is easier to take advantage of Theorem 1, which we have already proved.

The native has asserted (~C&B~k), and so the reasoner believes k≡(~C&B~k). But we know k≡(~C&B~k) is logically equivalent to ~k≡~(~C&B~k), which in turn is logically equivalent to ~k≡(B~k⊃C). Therefore the reasoner believes ~k≡(B~k⊃C), and so he believes a proposition of the form p≡(Bp⊃C)—p is the proposition ~k—and so, by Theorem 1, if he believes BC⊃C, he will believe C.

The solutions of Problems 2° and 3° can likewise be obtained as corollaries of Theorems 2 and 3, respectively. We leave the verification to the reader.

Exercise 5. Suppose a native says to a reasoner of type 4: "If you ever believe I'm a knight, then you will be inconsistent." Is it possible for the reasoner to believe in his own consistency without becoming inconsistent? (Hint: Use Theorem 2.)

Exercise 6. Suppose a reasoner of type 4 believes that if he believes the cure will work, then it will work. Suppose we now have a native who says to him: "If you ever believe that you will believe I'm a knight, then the cure will work."

Will the reasoner necessarily believe that the cure will work?

Exercise 7. Suppose the native instead says: "You will believe that if you ever believe I'm a knight, then the cure will work."

Will the reasoner necessarily believe that the cure will work?

Exercise 8. The following dialogue ensues between a student and his theology professor:

STUDENT: If I believe that God exists, then will I also believe that I will be saved?
PROFESSOR: If that is true, then God exists.
STUDENT: If I believe that God exists, then *will* I be saved?
PROFESSOR: If God exists, then that is true.

Prove that if the professor is accurate and if the student believes the professor, then God must exist and the student will be saved.

Exercise 9. The following strengthening of Theorem 3 can be proved. A reasoner of type 4 comes to the island for the cure and has the prior belief that if he should ever believe that the cure will work, then it will. He asks a native: "Will I ever believe that if you are a knight, then the cure will work?" The native replies: "If that is not so, then the cure will work." (Alternatively, he could have replied: "Either that is so or the cure will work.") The problem is to prove that the reasoner will believe that the cure will work. (Stated more abstractly, if a reasoner of type 4 believes $k \equiv (C \vee B(k \supset C))$ and believes $BC \supset C$, then he will believe C.) The proof of this is facilitated by first proving the following two facts as lemmas:

(1) For any propositions p and q, suppose a native says to a reasoner of type 4: "Either p is true or you will believe q." Then the reasoner will believe: "If the native is a knight, then I will believe that either p is true, or that the native is a knight."

(2) For any propositions p and q, suppose a native says to a reasoner of type 4: "Either p is true or else you will believe that if I am a knight, then q is true." The reasoner will then believe: "If the native is a knight, then either p is true or I will believe that q is true."

Exercise 10. The last exercise has the following dual. Again, a reasoner of type 4 believes that if he should believe that the cure will work, then it will. He now meets a native who says: "The cure doesn't work and you will never believe that either I'm a knight or that the cure works." Prove that the reasoner will believe that the cure will work.

• 16 •

The Rajah's Diamond

THOSE OF you who have read the magnificent story "The Rajah's Diamond," by Robert Louis Stevenson, will recall that at the end, the diamond was thrown into the Thames. Recent research, however, has revealed that the diamond was subsequently found by an inhabitant of the Island of Knights and Knaves who was vacationing in England at the time. One rumor has it that he then took the diamond to Paris and died shortly after. According to another, he took the diamond back home. If the second version is correct, then the diamond is somewhere on the knight-knave island.

A reasoner of type 4 decided to follow up on the second rumor in hopes of finding the diamond. He reached the island and believed the rules of the island. Also, the rules of the island really held. There are five different versions of what actually happened; each is of interest, and so I will relate them all.

1 · The First Version

According to the first version, when the reasoner reached the island, he met a native who made the following two statements:

(1) "If you ever believe that I am a knight, then you will believe that the diamond is on this island."

(2) "If you ever believe that I am a knight, then the diamond *is* on this island."

If this version is the correct one, is the diamond on the island?

2 · The Second Version

According to a second, slightly different version, the native, instead of making the two statements reported above, made the following two statements:

(1) "If I am a knave and if you ever believe that I'm a knight, then you will believe that the diamond is on this island."

(2) "I am actually a knight, and if you ever believe this, then the diamond is on this island."

If this second version is correct, is there sufficient evidence to conclude that the diamond must be on the island?

3 · The Third Version

The third version is particularly curious. According to it, the native made the following two statements:

(1) "If you ever believe that I'm a knight, then the diamond is not on this island."

(2) "If you ever believe that the diamond is on this island, then you will become inconsistent."

If this third version is correct, what conclusion should be drawn?

4 · The Fourth Version

According to the fourth version, the native made only one statement:

(1) "If you ever believe I'm a knight, then you will believe that the diamond is on this island."

The reasoner, of course, could get nowhere. He then discussed the whole affair with the Island Sage, who was known to be a knight

of the highest integrity. The Sage made the following statement:

"If the native you spoke to is a knight and if you ever believe that the diamond is on this island, then the diamond is on this island."

If this fourth version is correct, what conclusion should be drawn?

5 · The Fifth Version

This is like the last version, except that the native said:

(1) "You will believe that if I am a knight, then the diamond is on this island."

The Sage made the same statement.

Assuming that these five versions are equally probable, what is the probability that the Rajah's diamond is on the Island of Knights and Knaves?

Remarks. The mathematical content of the last two problems constitutes strengthenings of Theorems 2 and 3 of the last chapter—see the discussion following the solutions.

SOLUTIONS

We let k be the proposition that the native is a knight. We let D be the proposition that the diamond is on the island.

1 · Since the native made the two statements which he made, then the reasoner will believe the following two propositions:

(1) $k \equiv (Bk \supset BD)$
(2) $k \equiv (Bk \supset D)$

And the reasoner will therefore certainly believe the following two weaker propositions:

(1)′ $(Bk \supset BD) \supset k$
(2)′ $k \supset (Bk \supset D)$

As we will see, the fact that the reasoner believes even (1)′ and (2)′ is enough to solve the problem.

Since he believes (2)′, then by Lemma 1 of the last chapter, he will believe Bk⊃BD. Believing this and believing (1)′, he will then believe k. Believing k and believing (2)′, he will then believe Bk⊃D. Also, since he believes k, he will believe Bk, and hence he will believe D. Therefore BD is true. Hence Bk⊃BD is true, and since (1) is true (the rules of the island really hold), then k is true (the native is really a knight). Then since (2) is true, the proposition Bk⊃D is true. Also Bk is true (we have seen that the reasoner will believe k), and thus D is true. Therefore the diamond is on the island.

2 · According to this version, it does not follow from the native's two statements that the reasoner will believe the propositions (1) and (2) of the solution to the last problem, but it does follow that he will believe the weaker propositions (1)′ and (2)′ (see the note below). But, as we have seen in the solution of the last problem, this is enough to guarantee that the diamond is on the island.

Note: If a native of a knight-knave island says: "If I am a knave, then X," it logically follows that if X is true, the native must be a knight (because if X is true, then *any* proposition implies X, hence it is true that if the native is a knave, then X, but a knave couldn't make such a true statement). This is why the reasoner (who believes the rules of the island) will believe (1)′. As for (2)′, it is obvious that if a native says, "I am a knight and X," it follows that if the native is a knight, then X must be true.

3 · *Step 1:* The reasoner believes k≡(Bk⊃~D)—by virtue of the native's first statement. Then by Lemma 1 of the last chapter, the reasoner will believe Bk⊃B~D. And so the reasoner reasons: "If I ever believe that he is a knight, then I will believe that the diamond is not on the island. If I should also believe that the diamond *is* on

the island, then I will be inconsistent. Therefore, if I should ever believe he is a knight, then his second statement is true. And, of course, if his second statement is true, then he is a knight. This proves that if I should ever believe that he is a knight, then he really is a knight."

Step 2: The reasoner continues: "So suppose I believe he's a knight. Then he really is a knight, as I have just shown, hence his first statement Bk⊃~D is true. Also, if I believe he's a knight, then Bk is true, and thus ~D is true. Therefore, if I believe he's a knight, then the diamond is not on the island. He said just this in his first statement, and so he is a knight."

Step 3: The reasoner continues: "Now I believe he is a knight and I have already shown that Bk⊃~D, hence ~D is true—the diamond is not on this island."

Step 4: The reasoner now believes that the diamond is not on the island. Therefore, if he should ever believe that the diamond *is* on the island, he will be inconsistent. This proves that the native's second statement was true and therefore the native is in fact a knight. And so the native's first statement was also true—Bk⊃~D is true. But Bk is true (as we have proved), and so ~D is true. Therefore the diamond is not on the island.

4 • *Step 1:* The native asserted Bk⊃BD, and so the reasoner believes k≡(Bk⊃BD). Then by Lemma 1 of the last chapter (taking BD for p), the reasoner will believe Bk⊃BBD, and so the reasoner reasons: "Suppose I ever believe that the native is a knight. Then I will believe BD. Then I will believe k and I will believe BD, hence I will believe k&BD. I also believe the Sage's statement (k&BD)⊃D, so if I ever believe k&BD, then I will believe D. Therefore, if I ever believe k, I will also believe D—Bk⊃BD is true. This is what the native said, hence he is a knight."

Step 2: The reasoner continues: "I now believe k—Bk is true. Also Bk⊃BD is true (as I have proved), hence BD is true (I will *believe*

the diamond is on the island). Thus k is true and BD is true, so k&BD is true. Then, by the Sage's statement, D must be true—the diamond is on this island."

Step 3: The reasoner now believes D, so BD is true. Then Bk⊃BD is certainly true, hence the native is really a knight. Thus k and BD are both true, so k&BD is true. Hence, by the Sage's statement, D must be true—the diamond is on the island.

5 • The solution to this problem is somewhat simpler than the solution to Problem 4.

Step 1: Since the native said what he did, then by Lemma 3 of the last chapter, the reasoner will believe k⊃BD. Hence he will believe k⊃(k&BD). Since he also believes the Sage's statement (k&BD)⊃D, then he will believe k⊃D. Since the native said that he will believe k⊃D, then the reasoner will believe that the native is a knight—he will believe k. And since he will believe k⊃D, then he will believe D.

Step 2: Since the reasoner will believe k⊃D, then the native is really a knight. Hence k is true and also BD is true (as we have shown). Hence k&BD is true, and since (k&BD)⊃D is true, then D is true. So again the diamond is on the island.

Now we know that the chances are 80 percent that the diamond is on the island, since it is on the island in four out of five equally probable versions. This probability should be high enough to interest the enterprising reader who wishes to search for it.

Discussion. The essential mathematical content of Problem 4 is that for any propositions k and p, if a reasoner of type 4 believes k≡(Bk⊃Bp) and believes (k&Bp)⊃p, then he will believe p. This is a strengthening of Theorem 2 of the last chapter, because if a reasoner of type 4 believes Bp⊃p, he will certainly believe (k&Bp)⊃p, and so the present hypothesis is weaker than that of Theorem 2, Chapter 15 (and we have derived the same conclusion from a weaker assumption).

Likewise, the essential mathematical content of Problem 5 is that if a reasoner of type 4 believes $k \equiv B(k \supset p)$ and $(k \& Bp) \supset p$, then he will believe p. This is stronger than Theorem 3 of the last chapter for the same reasons.

Curiously enough, Theorem 1 of the last chapter does not appear to have an analogous strengthening. If a reasoner of type 4 believes $k \equiv (Bk \supset p)$ and $(k \& Bp) \supset p$, there does not seem to be any reason to conclude that he will believe p.

• 17 •

Löb's Island

SOMEWHERE IN the vast reaches of the ocean there is a particularly interesting knight-knave island which I shall refer to as Löb's Island. Given any person who visits the island, and given any proposition p, there is at least one native of the island who says to the visitor: "If you ever believe that I am a knight, then p is true."

In the problems of this chapter, a reasoner of type 4 visits the island. It is given that the rules of the island hold (knights tell the truth, knaves lie, and every native is a knight or a knave) and that the reasoner believes the rules of the island.

HENKIN'S PROBLEM

Suppose a native of Löb's Island says to the reasoner: "You will believe that I am a knight." On the surface it seems that there is no way to tell whether the native is a knight or a knave; it would appear that maybe the reasoner will believe that the native is a knight (in which case the native made a true statement, and therefore is a knight), or that maybe the reasoner will never believe that the native is a knight (in which case the native made a false statement and is accordingly a knave). Is there any way to decide between these two alternatives?

This problem derives from a famous problem posed by Leon Henkin and answered by M. H. Löb. The surprising thing is that it *is* possible to decide whether the native is a knight or a knave.

1

Under the conditions given above, is the native a knight or a knave? (Solutions are given following Problem 3.)

2

Suppose a native of Löb's Island says to a reasoner of type 4: "You will never believe that I am a knave." Assuming the rules of the island hold (and that the reasoner believes them), is the native a knight or a knave?

3

If a reasoner of type 4 visits Löb's Island (and believes the rules of the island), is it possible for him to be consistent and to believe that he is consistent?

SOLUTIONS TO PROBLEMS 1, 2, AND 3

1 · The solution is a bit tricky. Let P_1 be the native who said: "You will believe I'm a knight." Let k_1 be the proposition that P_1 is a knight. Since the reasoner believes the rules of the island and P_1 said what he did, then the reasoner will believe the proposition $k_1 \equiv Bk_1$. From just this fact, it is *not* possible to determine whether k_1 is true or false; but this island is Löb's Island, hence there is a native P_2 who will say to the reasoner: "If you ever believe I'm a knight, then P_1 is a knight." (Remember that for *any* proposition p, there is some native who says to the reasoner: "If you ever believe I'm a knight, then p.") Let k_2 be the proposition that P_2 is a knight. Since P_2

asserted the proposition $Bk_2 \supset k_1$, then the reasoner will believe the proposition $k_2 \equiv (Bk_2 \supset k_1)$. He also believes $Bk_1 \equiv k_1$, hence he believes $Bk_1 \supset k_1$, and so he believes $Bk_1 \supset k_1$ and $k_2 \equiv (Bk_2 \supset k_1)$. Then by Theorem 1, Chapter 15, page 126 (reading "k_2" for "k" and "k_1" for "C"), he will believe k_1. Since P_1 said that the reasoner would believe k_1, then P_1 is a knight. And so P_1 is a knight and the reasoner will believe that P_1 is a knight.

Note: We might remark that even without the assumption that the rules of the island really hold, we can still conclude that the reasoner will *believe* that P_1 is a knight.

2 · To avoid repetition of similar arguments, let us now note once and for all that if a reasoner of type 4 visits Löb's Island, then for any proposition p, if he believes the proposition $Bp \supset p$, he will believe p. (Reason: Since this island is Löb's Island, some native will say to him: "If you ever believe that I'm a knight, then p." Hence the reasoner will believe $k \equiv (Bk \supset p)$, where k is the proposition that the native is a knight. Then, since he believes $Bp \supset p$, he will believe p, by Theorem 1, Chapter 15.)

Now for the problem at hand: The native has told the reasoner, "You will never believe I'm a knave." Then the reasoner believes $k \equiv \sim B \sim k$ (k is the proposition that the native is a knight). Hence he will believe the logically equivalent proposition $\sim k \equiv B \sim k$, hence he will believe $B \sim k \supset \sim k$. Therefore he will believe $Bp \supset p$, when p is the proposition $\sim k$. Then, by the remarks of the last paragraph, he will believe p—i.e., he will believe $\sim k$. And so the reasoner will believe that the native is a knave. Since the native said that the reasoner wouldn't believe that the native is a knave, then the native in fact is a knave.

3 · Suppose a reasoner of type 4 visits Löb's Island. Then for any proposition p, there is some native who will say to him: "If you ever believe I'm a knight, then p." In particular, this is true if we take

for p the proposition $\perp$ (which, we recall, stands for logical falsehood). So there is a native who says to the reasoner: "If you ever believe I'm a knight, then $\perp$." Thus the reasoner believes the proposition $k \equiv (Bk \supset \perp)$. Now, $Bk \supset \perp$ is logically equivalent to $\sim Bk$, hence $k \equiv (Bk \supset \perp)$ is logically equivalent to $k \equiv \sim Bk$, hence the reasoner believes $k \equiv \sim Bk$. Then, according to Theorem 1, Chapter 12, page 101, he cannot believe in his own consistency without becoming inconsistent.

REFLEXIVITY

Reflexive Reasoners. We shall say that a reasoner is *reflexive* if for every proposition q, there is at least one proposition p such that the reasoner will believe $p \equiv (Bp \supset q)$.

Any reasoner who visits Löb's Island (and believes the rules of the island) will automatically become a reflexive reasoner, since for any proposition q, there is at least one native who will say: "If you ever believe I'm a knight, then q is true," and so the reasoner will believe $k \equiv (Bk \supset q)$, where k is the proposition that the native is a knight. However, a reasoner who has never visited Löb's Island might be a reflexive reasoner for completely different reasons (some of which we will consider in Chapter 25).

Let us note that if a *reflexive* reasoner of type 4 visits an *ordinary* knight-knave island (it doesn't have to be Löb's Island) and meets a native who says to him, "You will believe that I'm a knight," then the reasoner *will* believe that the native is a knight. (He doesn't need a second native to tell him, "If I'm a knight, then so is the first native.") Also, if a *reflexive* reasoner of type 4 goes to an *ordinary* knight-knave island and is told by a native, "You will never believe I'm a knave," the reasoner will believe that the native is a knave (as in Problem 2).

Also, no consistent reflexive reasoner of type 4 can believe he is consistent.

And, one more thing: Suppose a *reflexive* reasoner of type 4 is thinking of visiting the knight-knave island with the sulfur baths and mineral waters, and his family doctor, whom he trusts, tells him: "If you believe that the cure will work, then it will." Then without further ado, the reasoner will believe that the cure will work (he doesn't have to go first to the island and meet a native who tells him, "If you believe I'm a knight, then the cure will work").

All these facts are special cases of the following theorem (which springs from Theorem 1 of Chapter 15).

Theorem A (After Löb). For any proposition q, if a reflexive reasoner of type 4 believes $Bq \supset q$, he will believe q.

Reflexive Systems. Let us now consider the type of mathematical systems described in Chapter 13. We recall that for any system S, for any proposition p expressible in the system, the proposition Bp (p is provable in S) is also expressible in the system. (Remember that for systems, "B" means "provable.") When we have only one system S under discussion, the word "proposition" shall be understood to mean "proposition expressible in S."

We now define S to be *reflexive* if for every proposition q (expressible in S), there is at least one proposition p (expressible in S) such that the proposition $p \equiv (Bp \supset q)$ is *provable* in S. Theorem A above obviously holds for *systems* as well as reasoners: Given any reflexive system S and any expressible proposition q, if $Bq \supset q$ is provable in S, so is q. This is Löb's Theorem.

We shall call S a *Löbian* system if for every proposition p, if $Bp \supset p$ is provable in the system, so is p. We have now established Löb's Theorem.

Theorem L (Löb's Theorem). Every reflexive system of type 4 is Löbian—i.e., for any reflexive system S of type 4, if $Bp \supset p$ is provable in S, so is p.

Corollary. For any reflexive system of type 4, if p≡Bp is provable in the system, so is p.

Discussion. Gödel proved his incompleteness theorems for several systems, including the system of Arithmetic, which we have briefly mentioned. These systems are all reflexive systems of type 4, and this is what enabled Gödel's arguments to go through. Gödel constructed a sentence g that asserted its own nonprovability in the system (the sentence g≡~Bg is provable in the system).

Later the logician Leon Henkin constructed a sentence h such that h≡Bh is provable in the system, and raised the problem whether there was any way to tell whether h was provable in the system or not. Such a sentence h can be thought of as asserting: "I *am* provable in the system." (It resembles a native who says: "You *will* believe that I'm a knight.") On the surface, it would appear equally possible that h is true and provable in the system, or false and not provable in the system. The problem remained open for several years, and was finally solved by Löb. We have the answer in the corollary above: If the system is reflexive and of type 4, then Henkin's sentence h *is* provable in the system.

Reflexive and Gödelian Systems. We recall that a system is called *Gödelian* if there is some proposition p such that p≡~Bp is provable in the system.

4

Every reflexive system of type 1 is also Gödelian. Why is this? (Solutions are given following Problem 5.)

Strong Reflexivity. We will say that a system S is *strongly* reflexive if for every proposition q, there is a proposition p such that p≡B(p⊃q) is provable in S.

The connections between reflexivity and strong reflexivity are given in the following theorem.

Theorem R (Reflexivity Theorem). It consists of two parts:

(a) Any strongly reflexive system of type 1 is reflexive.

(b) Any reflexive system of type 1* is strongly reflexive.

5

Prove Theorem R.

SOLUTIONS TO PROBLEMS 4 AND 5

4 · Suppose S is reflexive and of type 1. Then for any proposition q, there is some p such that $p \equiv (Bp \supset q)$ is provable in S. We take for q the proposition $\perp$, and so there is some p such that $p \equiv (Bp \supset \perp)$ is provable in S. But $Bp \supset \perp$ is logically equivalent to $\sim Bp$, hence $p \equiv (Bp \supset \perp)$ is logically equivalent to $p \equiv \sim Bp$, and so $p \equiv \sim Bp$ is provable in S, and therefore S is Gödelian. (Actually, since we are basing propositional logic on $\supset$ and $\perp$, the proposition $\sim Bp$ *is* the proposition $Bp \supset \perp$, and so we really didn't even have to assume that S is of type 1.)

5 · Suppose S is of type 1. Let q be any proposition.

(a) Suppose S is strongly reflexive. Then there is a proposition p such that $p \equiv B(p \supset q)$ is provable in S. Since S is of type 1, it follows that $(p \supset q) \equiv (B(p \supset q) \supset q)$ is provable in S. (For any propositions X, Y, and Z, the proposition $(X \supset Z) \equiv (Y \supset Z)$ is a logical consequence of $X \equiv Y$. Taking p for X, $B(p \supset q)$ for Y, and q for Z, the proposition $(p \supset q) \equiv (B(p \supset q) \supset q)$ is a logical consequence of $p \equiv B(p \supset q)$.) Therefore there is a proposition p'—namely, $p \supset q$—such that $p' \equiv (Bp' \supset q)$ is provable in S. Hence S is reflexive.

(b) Suppose also that S is regular (and hence of type 1*) and that S is reflexive. Then there is a proposition p such that $p \equiv (Bp \supset q)$ is

provable in S. Since S is regular, it follows that $Bp \equiv B(Bp \supset q)$ is provable, hence $(Bp \supset q) \equiv (B(Bp \supset q) \supset q)$ is provable. We now take p' to be the proposition $Bp \supset q$, and we see that $p' \equiv (Bp' \supset q)$ is provable. Since there is a proposition p' such that $p' \equiv (Bp' \supset q)$ is provable in S, S is strongly reflexive.

Remarks. It of course follows from the above theorem and Löb's Theorem that any strongly reflexive system of type 4 is Löbian. We proved this another way in Theorem 3, Chapter 15.

• *Part VII* •

IN DEEPER WATERS

• 18 •

Reasoners of Type G

MODEST REASONERS

We have called a reasoner *conceited* if for every proposition p, he believes $Bp \supset p$. Now, if a reasoner believes p, then there is nothing immodest about his believing $Bp \supset p$. (Indeed, if he believes p and is of type 1, he will also believe $q \supset p$ for *any* proposition q whatsoever, since $q \supset p$ is a logical consequence of p. In particular, he will believe $Bp \supset p$.)

We shall now call a reasoner *modest* if for every proposition p, he believes $Bp \supset p$ only if he believes p (in other words, if he believes $Bp \supset p$, then he believes p). In analogy with *systems,* we might also call a modest reasoner a *Löbian* reasoner.

Löb's Theorem states that every reflexive system of type 4 is Löbian. Stated in terms of reasoners, every reflexive reasoner of type 4 is modest.

Many results that can be proved for a given reasoner under the assumption that he is a reflexive reasoner of type 4 can be proved more swiftly from the assumption that he is a modest reasoner of type 4. For example, suppose a modest reasoner of type 4 (or even a modest reasoner of type 1) believes that he is consistent. Then he believes $\sim B\bot$. Hence he believes the logically equivalent proposition $B\bot \supset \bot$. Then, being modest, he must believe $\bot$, which means that he

is inconsistent! And so no modest reasoner—even of type 1—can *consistently* believe in his own consistency. (This, of course, doesn't mean that he necessarily *is* inconsistent; he might happen to be consistent, but if he is consistent—and a modest reasoner of type 1—then he cannot believe he is consistent.)

REASONERS OF TYPE G

Let us say that a reasoner is modest with respect to a given proposition p if it is the case that if he believes $Bp \supset p$, then he also believes p. A reasoner, then, is modest if he is modest with respect to every proposition p. We now say that a reasoner *believes* he is modest with respect to p if he believes the proposition $B(Bp \supset p) \supset Bp$—he believes that if he should ever believe that his belief in p implies p, then he will believe p. We now say that a reasoner believes he is modest if for every proposition p, he believes he is modest with respect to p—in other words, for every proposition p, he believes $B(Bp \supset p) \supset Bp$.

By a *reasoner* of type G is meant a reasoner of type 4 who believes he is modest (for every proposition p, he believes $B(Bp \supset p) \supset Bp$). By a *system* of type G is meant a system of type 4 such that for every proposition p, the proposition $B(Bp \supset p) \supset Bp$ is provable in the system.

A great deal of research has been going on in recent years about systems of type G. George Boolos has devoted an excellent book† to the subject, which we strongly recommend as a follow-up to this volume.

Several questions about reasoners of type G readily present themselves. If a reasoner of type G *believes* he is modest, is he necessarily modest? We will show shortly that the answer is yes; indeed, we will see that any reasoner of type 1* who believes he is modest must be modest. Now, what about a reasoner of type 4 who *is* modest. Does he necessarily believe he is modest? We will show that the answer

†*The Unprovability of Consistency* (Cambridge University Press, 1979).

is yes, and hence that any reasoner of type 4 is modest if and only if he believes he is modest—in other words if and only if he is of type G. Then we will show a surprising result discovered independently by the logicians Saul Kripke, D. H. J. de Jongh, and Giovanni Sambin—namely, that any reasoner of type 3 who believes he is modest must be of type 4 (and hence of type G).

Another question: We know that any reflexive reasoner of type 4 is modest (this is Löb's Theorem). Is a modest reasoner of type 4 necessarily reflexive? That is, given a modest reasoner of type 4, is it necessarily true that for any proposition q, there is some proposition p such that he believes $p \equiv (Bp \supset q)$? It is not difficult to prove that the answer is yes; we will do this in the next chapter. And so by the end of the next chapter we will have proved that for any reasoner, the following four conditions are equivalent:

(1) He is a modest reasoner of type 4.
(2) He is of type G (he is of type 4 and believes he is modest).
(3) He is of type 3 and believes he is modest.
(4) He is a reflexive reasoner of type 4.

MODESTY AND BELIEF IN ONE'S MODESTY

For any proposition p, let Mp be the proposition $B(Bp \supset p) \supset Bp$. Thus Mp is the proposition that the reasoner is modest with respect to p. We are about to show that any reasoner of type 4 who believes he is modest really is modest. More specifically, we will show the stronger result that for any proposition p, if he believes that he is modest with respect to p (without necessarily believing that he is modest with respect to any other propositions), then he really is modest with respect to p (in fact this holds even for normal reasoners of type 1). The converse of this stronger fact is not necessarily true

—i.e., if a reasoner of type 4 is modest with respect to p, then he does not necessarily believe that he is modest with respect to p. However, we will show that if a reasoner of type 4 is modest with respect to the proposition Mp, then he will believe that he is modest with respect to p (in other words, if he is modest with respect to Mp, then he will believe Mp). From this it will of course follow that if a reasoner of type 4 is modest (with respect to all propositions), then he will know that he is modest.

1

Why is it true that any normal reasoner of type 1 (and hence any reasoner of type 4) who believes he is modest with respect to p must really be modest with respect to p?

2

By virtue of the last problem, given any proposition p, the proposition $B(Mp) \supset Mp$ is *true* for any reasoner of type 4. Prove that any reasoner of type 4 *believes* the proposition $B(Mp) \supset Mp$. (He knows that if he should *believe* that he is modest with respect to p, then he really is modest with respect to p.)

3

Using the last problem, show that any reasoner of type 4 who is modest with respect to Mp will believe that he is modest with respect to p.

It of course follows from Problem 1 that any reasoner of type 4 who believes he is modest, really is modest. And it follows from Problem 3 that if a reasoner of type 4 is modest (with respect to every proposition), then he must believe he is modest (because for every proposition p, he is modest with respect to Mp, hence according to

Problem 3, he believes he is modest with respect to p). We have thus established Theorem M.

Theorem M. A reasoner of type 4 is modest if and only if he believes he is modest.

By virtue of Theorem M, a reasoner is of type G if and only if he is a modest reasoner of type 4. Stated in terms of systems rather than reasoners, we have Theorem M_1.

Theorem M_1. For a system of type 4, the following two conditions are equivalent:

(1) For any proposition p, if $Bp \supset p$ is provable in the system, so is p.

(2) For any proposition p, the proposition $B(Bp \supset p) \supset Bp$ is provable in the system.

Stated more briefly, a system of type 4 is Löbian if and only if it is of type G.

We proved in the last chapter that every reflexive system of type 4 is Löbian—this is Löb's Theorem. Combining this with Theorem M_1, we have the important Theorem M_2.

Theorem M_2. Every reflexive system of type 4 is of type G.

Another proof of Theorem M_2 will be given in the next chapter.

THE KRIPKE, DE JONGH, SAMBIN THEOREM

We now wish to prove that any reasoner of type 3 who believes that he is modest must also believe that he is normal—and hence must be of type 4, and therefore of type G.

We will actually prove more. Let us say that a reasoner is normal with respect to a proposition p if his believing p implies that he will also believe Bp. A reasoner, then, is normal if and only if he is normal

with respect to every proposition p. We will say that a reasoner believes that he is normal with respect to p if he believes the proposition $Bp \supset BBp$. (Of course if the reasoner *is* normal with respect to p, then the proposition $Bp \supset BBp$ is true.) We will say that a reasoner believes he is normal if for every proposition p, he believes that he is normal with respect to p. A reasoner of type 4, then, is a reasoner of type 3 who believes he is normal. The Kripke, de Jongh, Sambin Theorem states that every reasoner of type 3 who believes he is modest will also believe he is normal (and hence will be of type G). We will prove the stronger result that every reasoner of type 1* who believes he is modest will also believe he is normal. But first we will prove the more elementary result that every reasoner of type 1* who *is* modest is also normal (this result may be new). Then we will prove sharper versions of both these results.

We recall from Chapter 10 that we are using the notation Cp for (p&BP), where we read Cp as "the reasoner *correctly* believes p." The propositions $Cp \supset p$, $Cp \supset Bp$, $p \supset (Bp \equiv Cp)$ are, of course, all tautologies. We first need the following lemma:

Lemma 1. Any reasoner of type 1* believes the following propositions:

(a) $BCp \supset BBp$
(b) $p \supset (BCp \supset Cp)$

4

Why is Lemma 1 true?

5

Now show that any modest reasoner of type 1* must be normal. More specifically (and this is a hint!), show that for any proposition p, if a reasoner of type 1* is modest with respect to Cp, then he must be normal with respect to p.

6

Now show that any reasoner of type 1* who believes he is modest must also believe that he is normal. More specifically, show that for any proposition p, if a reasoner of type 1* believes that he is modest with respect to Cp, then he will believe that he is normal with respect to p.

Since every reasoner of type 3 is also of type 1* (as we proved in Chapter 11), we have now established Theorem M_3.

Theorem M_3 (Kripke, de Jongh, Sambin). Every reasoner of type 3 who believes he is modest is of type G.

Stated for systems rather than reasoners, Theorem M_3 states that for any system of type 3, if all propositions of the form $B(Bp \supset p) \supset Bp$ are provable in the system, then all propositions of the form $Bp \supset BBp$ are provable in the system, and hence the system must be of type G.

The following two exercises provide some curious alternative ways of characterizing reasoners of type G.

Exercise 1. Show that for any reasoner of type 4, the following two conditions are equivalent.

(a) The reasoner is of type G.

(b) And, for any propositions p and q, the reasoner believes $B(q \supset p) \supset (B(Bp \supset q) \supset Bp)$.

Exercise 2. Prove that for any reasoner of type 4, the following two conditions are equivalent.

(a) He is of type G.

(b) For any propositions p and q, if he believes $(Bp \& Bq) \supset p$, then he will believe $Bq \supset p$.

SOLUTIONS

1 · By hypothesis, the reasoner believes Mp—the proposition $B(Bp \supset p) \supset Bp$. We are to show that he is modest with respect to p—i.e., that if he believes $Bp \supset p$, then he believes p. So we assume he believes Bp, and we are to show that he believes (or will believe) p.

Since he believes $Bp \supset p$ (by assumption), he believes $B(Bp \supset p)$ (since he is normal). Also, he believes $B(Bp \supset p) \supset Bp$ (by hypothesis). Then, being of type 1, he believes, or will believe, Bp. He also believes $Bp \supset p$. Hence he believes, or will believe, p.

2 · Essentially, this follows from Problem 1 and the fact that a reasoner of type 4 "knows" that he is of type 4, hence knows how he can reason. In more detail, the reasoner reasons thus:

"Suppose I ever believe Mp—that is, suppose I ever believe $B(Bp \supset p) \supset Bp$. I am to show that I am modest with respect to p—i.e., that if I ever believe $Bp \supset p$, then I will believe p. So suppose I believe $Bp \supset p$; it remains to show that I will believe p.

"And so I will assume that I will believe $B(Bp \supset p) \supset Bp$ and that I will believe $Bp \supset p$. I must then show that I will believe p. Well, since I will believe $Bp \supset p$ (by my second assumption), then I will believe $B(Bp \supset p)$. Since I also believe $B(Bp \supset p) \supset Bp$ (by my first assumption), then I will believe Bp. Once I believe Bp and $Bp \supset p$, then I will believe p."

At this point the reasoner has derived Bp from the two assumptions B(Mp) and $B(Bp \supset p)$, he will believe $(B(Mp) \& B(Bp \supset p)) \supset Bp$, and the logically equivalent proposition $B(MP) \supset (B(Bp \supset p) \supset Bp)$, which is the proposition $B(Mp) \supset Mp$.

3 · If a reasoner believes $B(Mp) \supset Mp$ and is modest with respect to Mp, then he will of course believe Mp (because for any proposition q, if he believes $Bq \supset q$ and is modest with respect to q, he will believe

q—and so this is the case if q is the proposition Mp). Now a reasoner of type 4 *does* believe B(Mp)⊃Mp (as we showed in the last problem), and so if he is modest with respect to Mp, he will believe Mp (which means he will believe he is modest with respect to p).

4 · The reasoner is assumed to be of type 1*.

(a) Since he is of type 1, he believes the tautology Cp⊃Bp. Being regular, he will then believe BCp⊃BBp.

(b) Since he is of type 1, he believes the tautology Cp⊃p. Being regular, he then believes BCp⊃Bp. Being of type 1, he then believes (p&BCp)⊃(p&Bp), which is a logical consequence of the last proposition. Thus he believes (p&BCp)⊃Cp (because Cp is the proposition p&Bp) and hence he will believe the logically equivalent proposition p⊃(BCp⊃Cp).

5 · Suppose a reasoner of type 1* is modest with respect to Cp. We are to show that he is then normal with respect to p.

Suppose he believes p. He also believes p⊃(BCp⊃Cp), according to (b) of Lemma 1, hence he will believe BCp⊃Cp. Believing this and being modest with respect to Cp, he will believe Cp. He also believes the tautology Cp⊃Bp, and since he believes Cp, he will believe Bp.

This proves that if he believes p, he will believe Bp, hence he is normal with respect to p.

6 · Suppose a reasoner of type 1* believes he is modest with respect to Cp. By (b) of Lemma 1, he believes p⊃(BCp⊃Cp). Since he is regular, he then believes Bp⊃B(BCp⊃Cp). But he also believes B(BCp⊃Cp)⊃BCp, since he believes he is modest with respect to Cp. Believing these last two propositions, he will believe Bp⊃BCp. He also believes BCp⊃BBp, according to (a) of Lemma 1, and will therefore believe Bp⊃BBp—he will believe that he is normal with respect to p.

Solution to Exercise 1. (a) A reasoner of type G will reason: "Suppose $B(q \supset p)$ and $B(Bp \supset q)$. This means I'll believe $q \supset p$ and I'll believe $Bp \supset q$, hence I'll believe $Bp \supset p$, and since I am modest, I'll believe p. Thus $(B(q \supset p) \& B(Bp \supset q)) \supset Bp$, or, what is logically equivalent, $B(q \supset p) \supset (B(Bp \supset q) \supset Bp)$."

(b) Suppose that for any proposition p and q, a reasoner of type 4 believes $B(q \supset p) \supset (B(Bp \supset q) \supset Bp)$. Then this is also true if q and p are the same proposition, and so he believes $B(p \supset p) \supset (B(Bp \supset p) \supset p)$. He also believes $B(p \supset p)$—because he believes the tautology $p \supset p$, and being normal, he then believes $B(p \supset p)$—and hence he believes $B(Bp \supset p) \supset Bp$. Therefore the reasoner is of type G.

Solution to Exercise 2. (a) Suppose a reasoner of type 4 believes $(Bp \& Bq) \supset p$. Then he will reason: "$(Bp \& Bq) \supset p$. Hence $Bq \supset (Bp \supset p)$. I now believe $Bq \supset (Bp \supset p)$. Now, suppose Bq. Then I'll believe Bq, and since I believe $Bq \supset (Bp \supset p)$, I'll believe $Bp \supset p$. And therefore $Bq \supset B(Bp \supset p)$. Also $B(Bp \supset p) \supset Bp$, and so $Bq \supset Bp$. Thus $Bq \supset (Bp \& Bq)$. And since $(Bp \& Bq) \supset p$, then $Bq \supset p$."

(b) Suppose a reasoner of type 4 is such that for any proposition p and q, if he believes $(Bp \& Bq) \supset p$, then he believes $Bq \supset p$. We will show that he is modest (and hence of type G).

Suppose he believes $Bp \supset p$. Then for any proposition q, he certainly believes $(Bp \& Bq) \supset p$. Hence, for any proposition q, he believes $Bq \supset p$. Well, take any proposition q such that he believes q (for example, take $q = T$). Then he believes Bq, and believing $Bq \supset p$, he will believe p.

• 19 •

Modesty, Reflexivity, and Stability

MORE ON REASONERS OF TYPE G

1

There is something very interesting about a consistent reasoner of type G—or even a consistent modest reasoner of type 1*—namely, that there is *no* proposition p such that he can believe that he doesn't believe p! (He cannot believe $\sim Bp$!) Why is this?

2

It hence follows that if a reasoner of type G believes that he doesn't believe p, then he will be inconsistent (even though it may be true that he doesn't believe p). Thus for any reasoner of type G, the proposition $B\sim Bp \supset B\bot$ is *true.*

Prove that for any reasoner of type G and any proposition p, the proposition $B\sim Bp \supset B\bot$ is not only true, but is *known* to be true by the reasoner.

3

This problem is a sharpening of an earlier theorem. Let us recall how we proved that every reflexive reasoner of type 4 is of type G. We did this in two stages: We first proved that every reflexive reasoner of type 4 is Löbian (Löb's Theorem), and then we proved that every Löbian reasoner of type 4 is of type G.

Now, let us consider a reasoner of type 4 who is *not* necessarily reflexive. It might be that for *some* proposition q, there is a proposition p such that the reasoner believes $p \equiv (Bp \supset q)$, and for some other proposition q, there is no such proposition p. This much we do know: If, for a given q, there is some p such that the reasoner believes $p \equiv (Bp \supset q)$, then if the reasoner believes $Bq \supset q$, he will also believe q (by Theorem 1, Chapter 15), and hence the proposition $B(Bq \supset q) \supset Bq$ is *true*. But does that mean that the reasoner necessarily *knows* that it is true? The answer is the solution to this problem:

Prove that for any propositions p and q, if a reasoner of type 4 believes $p \equiv (Bp \supset q)$, then he *believes* $B(Bq \supset q) \supset Bq$.

Discussion. Of course the solution to Problem 3 yields an alternative proof that any reflexive reasoner of type 4 must be of type G. I will now mention something else worth noting.

By virtue of Problem 3, given any propositions p and q, the proposition $B(p \equiv (Bp \supset q)) \supset (B(Bq \supset q) \supset Bq)$ is *true* for any reasoner of type 4. It can be shown that any reasoner of type 4 *knows* that the above proposition is true. The reader might try this as an exercise.

SOME FIXED-POINT PRINCIPLES†

We have now seen two different proofs that every reflexive reasoner of type 4 is of type G. We will shortly prove that every reasoner of

†These are special cases of a remarkable fact discussed in the final chapter.

type G is reflexive—for every q, there is some p such that he believes $p \equiv (Bp \supset q)$. But first some preliminary problems.

4

I have already warned you of the dangers of believing any proposition of the form $p \equiv \sim Bp$. If, however, you happen to be a reasoner of type G, then I'm afraid you have no alternative!

Given a reasoner of type G, find a proposition p such that he must believe $p \equiv \sim Bp$.

5

It is also true that given any reasoner of type G, there is a proposition p such that he believes $p \equiv B \sim p$. Prove this.

6

Given a proposition q, find a proposition p (expressible in terms of q) such that any reasoner of type G will believe $p \equiv B(p \supset q)$.

7

Do the same with $Bp \supset q$—i.e., find a proposition p such that any reasoner of type G will believe $p \equiv (Bp \supset q)$.

Note: The result of the last problem is that every reasoner of type G is reflexive. We have already proved that every reflexive reasoner of type 4 is of type G, and so we now have Theorem L*.

Theorem L.* A reasoner of type 4 is of type G if and only if he is reflexive.

SOME MORE FIXED-POINT PROPERTIES

8

Given a reasoner of type G and any propositions p and q, show that if the reasoner believes $p \equiv B(p \supset q)$, then he will believe $p \equiv Bq$.

9

Show that if a reasoner of type G believes $p \equiv (Bp \supset q)$, then he will believe $p \equiv (Bq \supset q)$.

10

A reasoner of type G goes to a knight-knave island (and believes the rules of the island), and asks a native whether he is married. The native replies: "You will believe that either I am a knight or I am married."

Will the reasoner necessarily believe that the native is married? Will he necessarily believe that he is a knight?

STABILITY

In preparation for the next chapter, we will now introduce the notion of *stability.*

We will call a reasoner *stable* if for every proposition p, if he believes that he believes p, then he really does believe p. We will call a reasoner *unstable* if he is not stable—i.e., if there is at least one proposition p such that the reasoner believes that he believes p,

but he does not actually believe p. Of course every accurate reasoner is automatically stable, but stability is a much weaker condition than accuracy. An unstable reasoner is inaccurate in a very strange way; indeed, instability is as strange a psychological characteristic as peculiarity.

We note that stability is the converse of normality. If a *normal* reasoner believes p, then he believes Bp, whereas if a *stable* reasoner believes Bp, then he believes p.

We shall use the terms *stable* and *unstable* for mathematical systems as well as reasoners. We will call a mathematical system S *stable* if for any proposition p, if Bp is provable in S, then so is p. We shall say that a reasoner is stable with respect to a particular proposition p if the proposition $BBp \supset Bp$ is true—i.e., if his belief that he believes p guarantees that he really does believe p. We shall say that he *believes* he is stable with respect to p if he believes the proposition $BBp \supset Bp$. Finally, we will say that he believes that he is stable if for every proposition p, he believes $BBp \supset Bp$ (for every p, he believes that he is stable with respect to p).

11

Prove that if a modest reasoner believes that he is stable, then he is either unstable or inconsistent.

Remark. The above result of course implies that no consistent stable reasoner of type G can ever know that he is stable.

SOLUTIONS

1 · Suppose a modest reasoner of type 1* believes $\sim Bp$. He also believes the tautology $\bot \supset p$, hence he believes $B\bot \supset Bp$ (because he is regular), hence he believes the logically equivalent proposition $\sim Bp \supset \sim B\bot$, hence he believes $\sim Bp \supset (B\bot \supset \bot)$. Since he believes $\sim Bp$,

he then believes ($B\bot\supset\bot$). Then, since he is modest, he will believe $\bot$, which means that he will be inconsistent. Therefore, if he is consistent (and a modest reasoner of type 1*), he will never believe $\sim Bp$.

2 · Any reasoner of type 4 (or even any normal reasoner of type 1*) will successively believe the following propositions:

(1) $\bot\supset p$
(2) $B\bot\supset Bp$
(3) $\sim Bp\supset\sim B\bot$
(4) $\sim B\bot\supset(B\bot\supset\bot)$—this is a tautology
(5) $\sim Bp\supset(B\bot\supset\bot)$—by (3) and (4)
(6) $B\sim Bp\supset B(B\bot\supset\bot)$

If the reasoner is of type G, he will also believe $B(B\bot\supset\bot)\supset B\bot$, hence he will believe $B\sim Bp\supset B\bot$.

3 · Suppose a reasoner of type 4 believes $p\equiv(Bp\supset q)$. We showed in Lemma 1, Chapter 15, page 125, that he will then believe $Bp\supset Bq$. Since he is normal, he will believe that he believes $p\equiv(Bp\supset q)$ and he will believe that he believes $Bp\supset Bq$. And so he reasons: "I believe $p\equiv(Bp\supset q)$, and I believe $Bp\supset Bq$. Now, suppose I ever believe $Bq\supset q$. Then, since I believe $Bp\supset Bq$, I will believe $Bp\supset q$. And, since I believe $p\equiv(Bp\supset q)$, I will believe p. Then I will believe Bp, and since I believe $Bp\supset q$, I will believe q. This shows that if I ever believe $Bq\supset q$, I will believe q."

At this point the reasoner believes $B(Bq\supset q)\supset Bq$.

4 and 5 · We will first solve Problem 5. Take p to be $B\bot$. We claim that any reasoner of type G believes $B\bot\equiv B\sim B\bot$ (and hence believes $p\equiv B\sim p$, where p is the proposition $B\bot$).

We showed in Problem 2 that for *any* proposition p, a reasoner of type G believes $B\sim Bp\supset B\bot$, hence (taking $\bot$ for p) he believes $B\sim B\bot\supset B\bot$. Also, he believes the tautology $\bot\supset\sim B\bot$, so he believes

$B\bot \supset B{\sim}B\bot$. And since he believes $B{\sim}B\bot \supset B\bot$, he must believe $B\bot \equiv B{\sim}B\bot$.

Now for the solution of Problem 4. We have just shown that the reasoner believes $B\bot \equiv B{\sim}B\bot$, hence he believes ${\sim}B\bot \equiv {\sim}B{\sim}B\bot$. And so he believes $p \equiv {\sim}Bp$, when p is now the proposition ${\sim}B\bot$.

Translated into words, a reasoner of type G believes the proposition that he is consistent if and only if he doesn't believe he is consistent. He also believes the proposition that he is inconsistent if and only if he believes that he is consistent.

6 · A solution is to take p to be Bq. Let us verify that this works. The reasoner believes the tautology $q \supset (Bq \supset q)$, and since he is regular, he then believes $Bq \supset B(Bq \supset q)$. He also believes $B(Bq \supset q) \supset Bq$ (since he is of type G), hence he must believe $Bq \equiv B(Bq \supset q)$. Therefore he believes $p \equiv B(p \supset q)$, where p is the proposition Bq.

7 · We have shown that the reasoner believes $Bq \equiv B(Bq \supset q)$. Then, by propositional logic, he will believe $(Bq \supset q) \equiv B(Bq \supset q) \supset q$. And so he will believe $p \equiv (Bp \supset q)$, where p is now the proposition $Bq \supset q$. (Note: We have just duplicated the proof of Theorem R, Chapter 17, page 148. According to Problem 6, the reasoner is strongly reflexive, and so by (a) of Theorem R, he is reflexive.)

8 · Suppose he believes $p \equiv B(p \supset q)$. Then by Lemma 3, Chapter 15, page 129, he will believe $p \supset Bq$. He also believes the tautology $q \supset (p \supset q)$, and being regular, he then believes $Bq \supset B(p \supset q)$. Also, since he believes $p \equiv B(p \supset q)$, he must believe $B(p \supset q) \supset p$. Since he believes $Bq \supset B(p \supset q)$ and $B(p \supset q) \supset p$, he will believe $Bq \supset p$. And so he believes $Bq \supset p$ and $p \supset Bq$ (as we have shown), hence he must believe $p \equiv Bq$.

9 · Suppose he believes $p \equiv (Bp \supset q)$. Then he will believe $p \supset (Bp \supset q)$, and also, by Lemma 1, Chapter 15, he will believe $Bp \supset Bq$. He believes the tautology $q \supset (Bp \supset q)$, hence he will believe $q \supset p$ (since he also believes $p \equiv (Bp \supset q)$). But he is regular, hence he will then

believe $Bq \supset Bp$. Believing this, together with $Bp \supset Bq$, he will believe $Bp \equiv Bq$. Then, using propositional logic, he will believe $(Bp \supset q) \equiv (Bq \supset q)$. Then, since he believes $p \equiv (Bp \supset q)$, he will believe $p \equiv (Bq \supset q)$.

10 · The reasoner won't have any idea whether the native is married or not, but he will believe that the native is a knight. We can see this as follows.

Let m be the proposition that the native is married. The native has asserted B(kvm), where k is the proposition that the native is a knight. Hence the reasoner will believe $k \equiv B(kvm)$. Then he will certainly believe $k \supset B(kvm)$. He also believes the tautology $k \supset (kvm)$, hence he will believe $Bk \supset B(kvm)$, since he is regular. He also believes $B(kvm) \supset k$, since he believes $k \equiv B(k \supset m)$. Hence he will believe $Bk \supset k$. Then, being of type G, he will believe k.

11 · Suppose he believes that he is stable. Then for every proposition p, he believes $BBp \supset Bp$, hence he believes $BB\bot \supset B\bot$. If he is modest, he will then believe $B\bot$ (because for *any* proposition q, a modest reasoner who believes $Bq \supset q$ will believe q, and so this is true in particular if q is the proposition $B\bot$). Since he believes $B\bot$, then if he is stable, he will believe $\bot$ and thus be inconsistent. This proves that if he believes $BB\bot \supset B\bot$, he cannot be modest, stable, and consistent. Therefore, if he is modest, stable, and consistent, he can never believe $BB\bot \supset B\bot$, and so he cannot know that he is stable with respect to $\bot$.