LOGICIANS WHO REASON ABOUT THEMSELVES

Raymond M. Smullyan
Department of Philosophy
Indiana University
Bloomington, IN 47405

ABSTRACT

By treating belief as a modality and combining this with problems about constant truth tellers and constant liars (knights and knaves) we obtain some curious epistemic counterparts of undecidability results in metamathematics. Gödel's second theorem gets reflected in a logician who cannot believe in his own consistency without becoming inconsistent. Löb's theorem reflects itself in a variety of beliefs which of their own nature are necessarily self-fulfilling.

We shall consider some "epistemic" problems related to undecidability results in metamathematics.

Our action shall take place on an island in which each native is classified as either a *knight* or a *knave*. Knights make only true statements and knaves make only false ones. Any such island will be called a *knight-knave* island. On such an island, no native can claim to be a knave, since no knight would falsely claim to be a knave and no knave would correctly claim to be one.

Our two main characters are a logician L who visits the island and meets a native N who makes a statement to L.

FOREVER UNDECIDED

We will say that L is (always) *accurate* if he never believes any false proposition.

Problem 1

An accurate logician L visits the island and meets a native N who makes a certain statement. Once the native has made this statement, it becomes logically impossible for L to ever decide whether N is a knight or a knave (if L should ever decide either way, he will lose his accuracy). What statement could N make to ensure this?

Solution

One solution is that N says: "You will never believe that I am a knight." If L ever believes that N is a knight, this will falsify N's statement, making N a knave and hence making L inaccurate in believing that N is a knight. Therefore, since L is accurate, he will never believe that N is a knight. Hence N's statement was true, so N really is a knight. It further follows that N will never have the false belief that N is a knave. And so L must remain forever undecided as to whether N is a knight or a knave.

Discussion

The character N (as well as L) will be constant throughout this article. We shall let k be the proposition that N is a knight. Now, whenever N asserts a proposition p, the reality of the situation is that $k \equiv p$ is true (N is a knight if and only if p). For any proposition p, we shall let Bp be the proposition that L does or will believe p. The native

N has asserted $\sim Bk$ (you will never believe I'm a knight) and so $k \equiv \sim Bk$ is true. From this we concluded that k must be true, but L can never believe k (assuming that L is always accurate).

More generally, given *any* proposition p such that $p \equiv \sim Bp$ is true, if L is accurate, then p is true, but L will never believe p (nor will he believe $\sim p$).

There is a complete parallelism between logicians who *believe* propositions and mathematical systems that *prove* propositions. When dealing with the latter, we will let Bp be the proposition that p is *provable* in the system.* In the system dealt with by Gödel, there is a proposition p such that $p \equiv \sim Bp$ is true (even provable) in the system. Under the assumption that all provable propositions of the system are true (under a standard interpretation), it then follows (by the argument of Problem 1) that p, though true, is not provable in the system--nor is the false proposition $\sim p$.

AN ACCURACY PREDICAMENT

For the problems that now follow, we must say more about the logician's reasoning abilities: We will say that he is of *type 1* if he has a complete knowledge of propositional logic--i.e. he sooner or later believes every tautology (any proposition provable by truth-tables) and his beliefs (past, present and future) are closed under modus ponens--i.e. if he ever believes p and believes $p \supset q$ (p implies q) then he will (sooner or later) believe q.

Of course these assumptions are highly idealized, since there are infinitely many tautologies, but we can assume that the logician is immortal.

We henceforth assume that L is of type 1. From this it follows that given any propositions that L believes, he will sooner or later believe every proposition that can be derived from them by propositional logic. We shall also make the inessential assumption that if L can derive a conclusion q from a proposition p taken as a premise, he will then believe $p \supset q$. [This assumption adds nothing to the set of L's beliefs, but it makes many of the arguments shorter and more transparent.]

It is to be understood in all the problems that follow, that when L visits the knight-knave island that he *believes* it is a knight-knave island, and so when he hears N assert a proposition p, then L *believes* the proposition $k \equiv p$ (N is a knight if and only if p).

*More precisely, we have a formula Bew(x) (read: "x is the Gödel number of a provable sentence") and for any proposition (sentence) p we let Bp be the sentence Bew($\bar{n}$), where n is the Gödel number of p.

We shall say that L *believes* he is (always) accurate if for every proposition p, L believes $Bp \supset p$ (he believes: "If I should ever believe p, then p must be true.") A reasoner (logician) who believes he is always accurate might aptly be called *conceited*.

Problem 2

A reasoner L of type 1 visits the island and N says to him: "You will never believe that I'm a knight." The interesting thing now is that if L believes that he is always accurate, then he will become inaccurate. Why is this?

Solution

Suppose L believes that he is always accurate. Then he will reason: "If N is a knave, his statement is false, which means that I *will* believe he's a knight and hence be inaccurate. This is impossible, since I am always accurate. Therefore he can't be a knave; he must be a knight."

At this point L believes that N is a knight, which makes N's statement false, hence N is really a knave. Thus L is inaccurate in believing that N is a knight. [We might remark that if L hadn't assumed his own accuracy, he would never have lapsed into an inaccuracy. He has been justly punished for his conceit!]

Peculiarity

We will 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 strange condition doesn't necessarily involve a logical inconsistency, but it is certainly a psychological peculiarity!]

Problem 2A

Show that under the hypotheses of Problem 2, L will become not only inaccurate, but peculiar!

Solution

We have seen that L will believe that N is a knight. Then L will believe what N said and hence believe that he doesn't believe that N is a knight.

Remark. Even if the island that L visits is not a really a knight-knave island, but L only *believes* that it is (he believes $k\equiv\sim Bk$) the above argument goes through (though the argument of Problem 2 does not).

THE GÖDEL CONSISTENCY PREDICAMENT

We shall say that L is of *type 2* if he is of type 1 and also *knows* that his beliefs are closed under modus ponens--i.e. for every p and q he (correctly) believes: "If I should ever believe both p and $p\supset q$, then I will believe q." And so he believes $(Bp\&B(p\supset q))\supset Bq$ [and being of type 1, he also believes the logically equivalent proposition: $B(p\supset q)\supset(Bp\supset Bq)$].

We shall call a reasoner *normal* if whenever he believes p, he also believes that he believes p. We shall say that he *believes* he is normal if he believes all propositions of the form $Bp\supset BBp$ (he believes: "If I should ever believe p, then I will believe that I believe p.") We define a reasoner to be of *type 3* if he is a normal reasoner of type 2. If he also *believes* that he is normal, then we define him to be of type 4. Our main concern will be with reasoners of type 4. [They are the counterparts of mathematical systems of type 4--defined analogously, only reading "provable" for "B".]

A reasoner of type 4 (or even type 3) who believes $p\supset q$ will also believe $B(p\supset q)$ (by normality), hence will believe $Bp\supset Bq$ (since he believes $B(p\supset q)\supset(Bp\supset Bq)$). This means that if a reasoner L of type 4 visits a knight-knave island (or even one that he *believes* is a knight-knave island) and hears a native N assert a proposition p, then he will not only believe $k\supset p$ (which he will, since he will believe $k\equiv p$), but will also believe $Bk\supset Bp$ (he will believe: "If I should ever believe he's a knight, then I will believe what he said."). He will also believe $B\sim k\supset B\sim p$.

We shall define a reasoner to be *consistent* if he never believes any proposition and its negation. [An inconsistent reasoner of even type 1 will sooner or later believe *every* proposition q, since $p\supset(\sim p\supset q)$ is a tautology. Also, if we take some fixed contradictory proposition f (for logical *falsehood*), a type 1 reasoner is consistent if and only if he never believes f, since for every p, the proposition $f\supset p$ is logically true, hence if the reasoner believes f, he will believe p).

We will say that a reasoner *believes* he is consistent if for every proposition p he believes $\sim(Bp\&B\sim p)$ (he believes: "I will never believe both p and $\sim p$"). [For a reasoner of type 4--or even of type 3,--this is equivalent to his believing that he will never believe f. This is not hard to show.]

Now we come to our first "big" problem.

Problem 3

[After Gödel's Second Theorem] - A logician L of type 4 visits a knight-knave island (or at least he *believes* it to be one) and meets N who says: "You will never believe that I'm a knight." Prove that if L is consistent, he can never know that he is--or put another way, if L ever believes that he is consistent, he will become inconsistent.

Solution

Suppose L is confident of his consistency. Then he will reason: "Suppose I never believe he's a knight. Then I'll believe what he said--I'll believe that I don't believe he's a knight. But if I ever believe he's a knight, I'll also believe that I *do* believe he's a knight (since I am normal). This means I would be inconsistent, which isn't possible (sic!). Therefore, I never will believe he's a knight. He said I never would, hence he's a knight."

At this point, L believes that N is a knight. Being normal, he then continues: "Now I believe he's a knight. He said I never would, so he's a knave."

At this point L is inconsistent (a while ago he believed N was a knight).

Discussion

Isn't it possible for a consistent reasoner of type 4 to know that he is consistent? Yes, but only if he believes no proposition of the form $p \equiv \sim Bp$. [In the above problem, for example, L should have had the good sense to doubt that he was really on a knight-knave island!] However, with the type of mathematical system investigated by Gödel, the analogous option is not open--there really *is* a proposition p such that $p \equiv \sim Bp$ is provable in the system (and the system is of type 4). And so, by an analogous argument, the system, if consistent, cannot prove its own consistency (say in the form that $\sim Bf$ is not provable).

Henkin's Problem

For the same system, there is also a proposition p such that $p \equiv Bp$ is provable in the system. [This is like a native of the island who says: "You *will* believe that I'm a knight."] On the fact of it, p could be true and provable, or false and unprovable; is there any way to tell which? This problem remained open for some years and was finally solved by Löb, who showed the stronger fact that if $Bp \supset p$ is provable in the system, so is p. [His proof utilized the fact that there is also another proposition q such that $q\equiv(Bp\supset q)$ is provable in the system.] We now turn to a striking epistemic version of this.

SELF-FULFILLING BELIEFS AND LÖB'S THEOREM

We now have a change of scenario. A logician L of type 4 is *thinking* of visiting the island of knights and knaves because he has heard a rumor that the sulphur baths and mineral waters there might cure his rheumatism. He is home discussing this with his family physician and asks: "Does the cure really work?" 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."

The logician fully trusts his doctor and so he goes to the island with the prior belief that if he should believe that the cure will work, then it will. He takes the cure, which lasts a day, and which is supposed to work in a few weeks (if it works at all). But the next day, he starts worrying: He thinks: "I know that if I should believe that the cure will work, then it will, but what evidence do I have that I will ever *believe* that the cure works? And so how do I know that it will?"

A native N passes by and asks L why he looks so disconsolate. L explains the situation and concludes: "--and so how do I know that the cure will work?" N then draws himself up in a dignified manner and says: "If you ever believe I'm a knight, then the cure will work."

Problem 4

Amazingly enough, the logician *will* believe that the cure will work, and, if his doctor was right, it will. How is this proved?

Solution

We let c be the proposition that the cure will work. L has the prior belief that $Bc\supset c$. Also, since N said that $Bk\supset c$, L believes $k\equiv(Bk\supset c)$. And so L reasons: "Suppose I ever believe that he's a knight (suppose I believe k). Then I'll believe what he said--I'll believe $Bk\supset c$. But if I believe k, I'll also believe Bk (since I am normal). Once I believe Bk and believe $Bk\supset c$, I'll believe c. Thus, if I ever believe he's a knight, then I'll *believe* that the cure will work. But if I believe that the cure will work, *then* it will (as my doctor told me). Therefore, if I ever believe he's a knight, then the cure will work. Well, that's exactly what he said, hence he's a knight!"

At this point, L believes that N is a knight. Since L is normal, he continues: "Now I believe he's a knight. And I have already proved that if I believe he's a knight, then the cure will work. Therefore the cure will work."

The logician now believes that the cure will work. Then (if his doctor was right), it will.

Reflexive Reasoners (and Systems)

Generalizing the above problem, for any proposition p, if a reasoner of type 4 believes $Bp \supset p$, and if there is a proposition q such that he believes $q \equiv (Bq \supset q)$, then he will believe p.

We will call a reasoner reflexive if for every proposition p there is some q such that the reasoner believes $q \equiv (Bq \supset p)$. And so if a reflexive reasoner of type 4 believes $Bp \supset p$, he will believe p. This is Löb's theorem (for reasoners).

For systems, we define reflexivity to mean that for any p (in the language of the system) there is some q such that $q \equiv (Bq \supset p)$ is provable in the system. Löb's theorem (in a general form) is that for any reflexive system of type 4, if $Bp \supset p$ is provable in the system, so is p.

Remarks

Here are some variants of Problem 4 that the reader might like to try as exercises: Suppose N had instead said: "The cure doesn't work and you will believe that I'm a knave." Prove that L will believe that the cure works.

Here are some other things that N could have said to ensure that L will believe that the cure works:

(1) If you believe that I'm a knight, then you will believe that the cure will work.

(2) You will believe that if I am a knight then the cure will work.

(3) You will believe I'm a knave, but you will never believe that the cure will work.

(4) You will never believe either that I'm a knight or that the cure will work.

THE STABILITY PREDICAMENT

We will call a reasoner *unstable* if there is some proposition p such that he believes that he believes p, but doesn't really believe p. [This is just as strange a psychological phenomenon as peculiarity!]

We will call him *stable* if he is not unstable--i.e. for every p, if he believes Bp then he believes p. [Note that stability is the converse of normality.] We will say that a reasoner *believes* that he is stable if for every proposition p, he believes $BBp\supset Bp$ (he believes: "If I should ever believe that I believe p, then I really will believe p).

Problem 5

If a consistent reflexive reasoner of type 4 believes that he is stable, then he will become unstable. Stated otherwise, if a stable reflexive reasoner of type 4 believes that he is stable, then he will become inconsistent. Why is this?

Solution

Suppose that a stable reflexive reasoner of type 4 believes that he is stable. We will show that he will (sooner or later) believe every proposition p (and hence be inconsistent).

Take any proposition p. The reasoner believes $BBp\supset Bp$, hence by Löb's theorem he will believe Bp (because he believes $Br\supset r$, where r is the proposition Bp, and so he will believe r, which is the proposition Bp). Being stable, he will then believe p.

A QUESTION OF TIMIDITY

The following problem affords another (and rather simple) illustration of how a belief can be self-fulfilling.

Problem 6

A certain country is ruled by a tyrant who owns a brain-reading machine with which he can read the thoughts of all the inhabitants. Each inhabitant is a normal, stable reasoner of type 1.

There is one particular proposition E which all the inhabitants are *forbidden* to believe--any inhabitant who believes E gets executed! Now, given any proposition p, we will say that it is *dangerous* for a given inhabitant to believe p if his believing p will lead him to believing E.

The problem is to prove that for any proposition p, if a given inhabitant *believes* that it is dangerous for him to believe p, then it really is dangerous for him to believe p.

Solution

Suppose an inhabitant does believe that it is dangerous for him to believe p. He thus believes the proposition $Bp \supset BE$. We will show that $Bp \supset BE$ is therefore true--i.e. if he should ever believe p, then he really will believe E.

Suppose he believes p. Being normal, he will then believe Bp. And since he also believes $Bp \supset BE$ and is of type 1, he will believe BE. Then, since he is stable, he will believe E.

A GRAND INDECISION

We again consider a reflexive, stable reasoner of type 4. There is a proposition p such that he can never believe p and can never believe $\sim p$ without becoming inconsistent in either case. [And so if he is consistent, he will never believe either one.] Can you find such a proposition p? [Note: Unlike Problem 1, we are *not* assuming that the reasoner is always accurate.]

Solution

We let f be any tautologically contradictory proposition--any proposition such that $\sim f$ is a tautology. Then for any proposition q, the proposition $\sim q \equiv (q \supset f)$ is a tautology, and so any reasoner--even of type 1--who believes $\sim q$ will believe $q \supset f$.

We now take for p the proposition Bf--the proposition that the reasoner believes (or will believe) f. [Of course if a reasoner of type 1 believes f, he will be inconsistent, since $f \supset p$ is a tautology for every p].

If the reasoner should believe Bf, he will believe f (since he is stable) and hence will be inconsistent. On the other hand, if he should ever believe $\sim Bf$, he will believe $Bf \supset f$, and so by Löb's theorem, he will believe f and again be inconsistent. [This last observation is due to Georg Kreisel.]

MODEST REASONERS

We have called a reasoner *conceited* if he believes all propositions of the form $Bp \supset p$. At the other extreme, let us call a reasoner *modest* if he never believes $Bp \supset p$ unless he believes p. [If he believes p and is of type 1, he will, of course, *have* to believe $Bp \supset p$--in fact $q \supset p$ for any q whatsoever]. Löb's theorem (for reasoners) can be succinctly stated: Any reflexive reasoner of type 4 is modest.

The theory of modest reasoners of type 4 (or rather the analogous theory for systems) is today an elaborate one, of which we can say here but a little. For one thing, it can be shown that any modest reasoner of type 4 must be reflexive (a sort of converse of Löb's theorem). Another thing: Let us say that a reasoner *believes* he is modest if for every p, he believes the proposition $B(Bp \supset p) \supset Bp$. [Of course, all these propositions are *true* if the reasoner really is modest.] It is not difficult to show that any reasoner of type 4 (or even any normal reasoner of type 1) who believes he is modest really is modest. [The reader might try this as an exercise.] It can also be shown (but this is a bit more tricky) that every modest reasoner of type 4 believes that he is modest. A surprising result (due to Kripke, deJongh and Sambin) is that every reasoner of type 3 who believes he is modest will also believe he is normal--and thus is of type 4! And so for any reasoner, the following 4 conditions are equivalent: (1) He is a reflexive reasoner of type 4; (2) He is a modest reasoner of type 4; (3) He is a reasoner of type 4 who believes he is modest; (4) He is a reasoner of type 3 who believes he is modest. [Proofs of these equivalences can be found in [2], or in a more formal version, in [1].]

Reasoners satisfying any of the above equivalent conditions correspond to an important system of modal logic known as G--accordingly, they are called (in [2]) *reasoners of type G*. Boolos [1] has devoted an excellent book to this modal system and [2] contains a host of epistemic problems about reasoners of this and other types. [A particularly curious reasoner to be met in [2] is the *queer* reasoner--he is of type G and believes that he is inconsistent. But he is wrong in this belief!]

I wish to conclude with another epistemic puzzle which I think you might enjoy trying as an exercise.

A reasoner of type 4 (not necessarily reflexive) goes to an island which is and which he believes to be a knight-knave island. He visits it because of a rumor that there is gold buried there. He meets a native and asks: "Is there really gold here?" The native then makes two statements: (1) If you ever believe I'm a knight, then you will believe that there is gold here; (2) If you ever believe I'm a knight, then there *is* gold here.

Is there gold on this island or not? Why?

References

G. Boolos. The Unprovability of Consistency. Cambridge University Press (1979).

R. Smullyan. Forever Undecided: A Puzzle Guide to Gödel. Knopf (In Press).