Gödel's Second Incompleteness Theorem

Let Γ be a recursively axiomatized theory that includes Q. The proof of Gödel's first incompleteness theorem proceeded by constructing a sentence σ such that

(1) $(\sigma \leftrightarrow \neg\, \mathrm{Bew}_{\Gamma}([\ulcorner \sigma \urcorner]))$

is a theorem. Having this, it was straightforward to prove the following:

(2) If Γ is consistent, then σ isn't provable in Γ.

We can formulate this result withing the language of arithmetic. Using "Con(Γ)" to abbreviate "$\sim \mathrm{Bew}_{\Gamma}([\ulcorner \sim 0 - 0 \urcorner])$," we formalize (2) as:

(3) $(\mathrm{Con}(\Gamma) \rightarrow\, \sim \mathrm{Bew}_{\Gamma}([\ulcorner \sigma \urcorner]))$.

(1) and (3) together give us:

(4) $(\mathrm{Con}(\Gamma) \rightarrow \sigma)$.

Provided that Γ includes PA, we can formalize the derivation of (4) within Γ. Hence we have:

(5) $\mathrm{Bew}_{\Gamma}([\ulcorner(\mathrm{Con}(\Gamma) \rightarrow \sigma)\urcorner])$.

Because the set of theorems of Γ is closed under *modus ponens*, (5) gives us:

(6) $\{\mathrm{Bew}_{\Gamma}([\ulcorner \mathrm{Con}(\Gamma)\urcorner] \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner \sigma \urcorner]))$

(3) and (6) together tautologically imply:

(7) $(\mathrm{Con}(\Gamma) \rightarrow\, \sim \mathrm{Bew}_{\Gamma}([\ulcorner \mathrm{Con}(\Gamma)\urcorner]))$.

Thus we have:

> **Second Incompleteness Theorem.** No consistent, recursively axiomatized theory that includes PA can prove its own consistency.

Proof: I don't really want give a proof of (5), which would be unbearably long and boring. I'll be content with a sketch. The proof depends on three principles, which were first singled out by M. H. Löb:

(L1) If $\Gamma \vdash \phi$, then $\Gamma \vdash \mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner])$.

(L2) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner])\urcorner]))$.

(L3) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner(\phi \rightarrow \psi)\urcorner]) \rightarrow (\mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\psi\urcorner])))$.

(L1) we proved earlier, when we showed that "$\mathrm{Bew}_{\Gamma}(x)$" weakly represents $\{\ulcorner\phi\urcorner : \phi$ is a consequence of $\Gamma\}$ in Q. This tells us that, if $\Gamma \vdash \phi$, then $Q \vdash \mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner])$, and so, since Γ includes Q, $\Gamma \vdash \mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner])$.

(L2) is the formalized statement of (L1). To prove it, we formalize our proof that every true Σ sentence is provable in Γ. First we show, by induction on the complexity of terms, that, for each term $\tau(x_1,\ldots,x_n)$,

$$\Gamma \vdash (\forall z_1)\ldots(\forall z_n)(\forall z_{n+1})(\tau(z_1,\ldots,z_n) = z_{n+1} \rightarrow$$

$$\mathrm{Bew}_{\Gamma}([\ulcorner\tau(x_1,\ldots x_n) = x_{n+1}\urcorner{}^{x_1}/_{[z_1]}\ldots{}^{x_n}/_{[z_n]}{}^{x_{n+1}}/_{[z_{n+1}]}])).$$

Next show, by induction on the complexity of bounded formulas, that for each bounded formula $\psi(x_1,\ldots,x_n)$,

$$\Gamma \vdash (\forall z_1)\ldots(\forall z_n)(\psi(z_1,\ldots,z_n) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\psi(x_1,\ldots,x_n)\urcorner{}^{x_1}/_{[z_1]}\ldots{}^{x_n}/_{[z_n]}])).$$

Next, prove the same thing for ψ a Σ formula, by induction on the length of the initial existential quantifier prefix. The key fact we need is that our rules of proof include Existential Generalization. Because $\mathrm{Bew}_{\Gamma}(x_1)$ is a Σ formula, a special case of this principle will be:

$$\Gamma \vdash (\forall z_1)(\mathrm{Bew}_{\Gamma}(z_1) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\mathrm{Bew}_{\Gamma}(x_1)\urcorner{}^{x_1}/_{[z_1]}])).$$

We get (L2) by instantiating with $[\ulcorner\phi\urcorner]$. A detailed proof would be quite laborious.

To prove (L3), simply note that, if we have proofs of $(\phi \rightarrow \psi)$ and ϕ, we get a proof of ψ by taking the two proofs together, then adding ψ at the end by the rule Modus Ponens. All we need to know is that the arithmetical concatenation operation works properly.

Now that we have (L1)-(L3), we want to see how to utilize them to get the Second Incompleteness Theorem. The Self-Referential Lemma gives us a sentence σ such that

(a) $\Gamma \vdash (\sigma \leftrightarrow {\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]))$.

We have,

(b) $\Gamma \vdash (\sigma \rightarrow {\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])))$.

(L1) lets us derive:

(c) $\Gamma \vdash \mathrm{Bew}_{\Gamma}([\ulcorner(\sigma \rightarrow {\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]))\urcorner])$.

Applying (L3), we get:

(d) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner{\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])\urcorner]))$.

(L2) gives us this:

(e) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])\urcorner]))$

$({\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow (\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow {\sim}0 = 0))$ is a tautology, hence a theorem of Γ, so that by (L1) we have:

(f) $\Gamma \vdash \mathrm{Bew}_{\Gamma}([\ulcorner({\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow (\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow {\sim}0 = 0))\urcorner])$.

Two applications of (L3) give us:

(g) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner{\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])\urcorner]) \rightarrow (\mathrm{Bew}_{\Gamma}([\ulcorner\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])\urcorner]) \rightarrow$
$\mathrm{Bew}_{\Gamma}([\ulcorner{\sim}0 = 0\urcorner])))$.

(d), (e), and (g) yield, by truth-functional logic:

(h) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner{\sim}0 = 0\urcorner]))$.

(i) $\Gamma \vdash (\mathrm{Con}(\Gamma) \rightarrow {\sim}\mathrm{Bew}_{\Gamma}([\ulcorner\sigma\urcorner]))$,

using the definition of "Con." This and (a) give us:

(j) $\Gamma \vdash (\mathrm{Con}(\Gamma) \rightarrow \sigma)$.

We want to show that, if Γ is consistent, Con(Γ) isn't provable in Γ. Equivalently, we want to show that, if Con(Γ) is provable in Γ, then Γ is inconsistent. Toward this end, suppose that

(k) $\Gamma \vdash \text{Con}(\Gamma)$.

(j) and (k) yield:

(l) $\Gamma \vdash \sigma$.

By (L1), we get:

(m) $\Gamma \vdash \text{Bew}_{\Gamma}([\ulcorner\sigma\urcorner])$.

(a) and (m) give us:

(n) $\Gamma \vdash \neg\, \sigma$.

(l) and (n) show us that Γ is inconsistent, as required.⊠

The First Incompleteness Theorem was a little disappointing. Sure, it did what it was advertised as doing, giving us a true, unprovable sentence. But the sentence it gave us was an out-of-the-way statement that, apart from its appearance in the First Incompleteness Theorem, no one would ever have been interested in. The theorem left open the possibility that there aren't any interesting statements that aren't provable in PA.

With the Second Incompleteness Theorem, our disappointment dissipates, for we certainly are interested in knowing whether our theories are consistent. Thus Con(PA) is an important truth that isn't provable in PA.

We've proved the Second Incompleteness Theorem for the language of arithmetic, but the same proof will work for any language into which we can translate the language of arithmetic. In particular, we can prove that, if axiomatic set theory is consistent, then it cannot prove its own consistency. This discovery stopped Hilbert's program dead in its tracks. Hilbert hoped to prove the consistency of the axioms of set theory in a system much weaker than axiomatic set theory. It turns out that (assuming axiomatic set theory is consistent) proving its consistency will require a theory stronger than axiomatic set theory. So if you're worried about the consistency of set theory, you won't get a consistency proof that helps.

By using interpretations, we can extend the proof of the Second Incompleteness Theorem to theories like the Zermelo-Fraenkel axioms for set theory, theories that are substantially richer than PA. We can use the same technique to extend the proof to theories that are substantially weaker than PA, like Q. We cannot prove the Second Incompleteness Theorem for Q directly, by proving analogues to (L1)-(L3), with "Bew_Q" in place of Bew_{PA}." To prove (L2), we need induction, and in Q we don't have induction. To obtain the Second Incompleteness Theorem for Q, and for other theories that include Q, one has to be more devious. We first produce a theory Γ that, although weaker that PA – it's obtained from PA by restricting the induction axiom schema – is nonetheless strong enough to prove Second Incompleteness Theorem. Next, we interpret Γ into Q by giving a translation that leaves the nonlogical terms alone but suitably restricts the quantifiers. If we had a proof of CON(Q) in Q, we could get a proof of CON(Γ) in Γ. The details of the proof, which is mainly due to Alex Wilkie, are delicate, and all I can do here is give you the URL for Sam Buss's web site, where you can find the proof written out; it's http://www.math.ucsd.edu/~sbuss/ResearchWeb/handbookII/index.html.

Refocusing our attention on PA, if we think a sentence θ is true, we'll think it's consistent; that is, we would expect to have the following:

$$(\gamma \rightarrow \text{Con}(\{\gamma\}).$$

All instances of this schema are, in fact, provable in PA. The proof isn't easy, and I won't attempt it here. It follows that, if the theory Γ is finitely axiomatized, it proves its own consistency. Consequently, by the Second Incompleteness Theorem, if Γ is finitely axiomatized and it includes PA, then Γ is inconsistent. Thus we have the following:

> **Theorem** (Ryll-Nardzewski). No consistent theory in the language of arithmetic that includes PA can be finitely axiomatized.

This theorem is not as resilient as most of the results we have been studying, which generalize from the language of arithmetic to other languages into which you can translate the language of arithmetic. Ryll-Nardzewski's theorem, and the schema $(\gamma \rightarrow \text{Con}(\{\gamma\}))$ that backs it up, are brittle. They hold for the language of arithmetic, but they don't necessarily hold for other languages that include the language of arithmetic.

Let us say – this is a little vague, but bear with me– that we *consciously accept* an arithmetical theory Γ if, upon reflection, we are willing to agree that all the members of Γ are true. (Keep in mind what we learned from Tarski, that, whereas the general notion of truth is philosophically suspicious, truth in the language of arithmetic is unexceptionable.) If we think that all the members of Γ are true, then we must surely think that Γ is consistent, since we know the Soundness Theorem, which tells us that the members of an inconsistent set of sentences can't all be true. Thus if Γ is a recursively axiomatized set of sentences that includes PA, then if we consciously accept Γ, we'll accept Con(Γ).

Let Γ be the set of arithmetical sentences that we are, upon careful reflection, willing to accept. Assuming, *pace* Lucas and Penrose, that we Γ is effectively enumerable, that if we consciously accept Γ, then we'll accept Con(Γ), which means, by the Second Incompleteness

Theorem, that Γ is inconsistent. This puts a limitation on our capacity for self-knowledge. Assuming the set of arithmetical sentences we accept is consistent, we can't *consciously* accept the set of arithmetical sentences we accept. Perhaps we consciously accept each finite subset of the set of arithmetical sentences we accept, but we won't consciously accept the whole thing.

There are a couple of ways this could happen. The most likely scenario is that we aren't sufficiently aware of our own mental state to be able to identify the set of arithmetical sentences we accept. This is, it seems to me, the normal human condition.

We can, however, imagine science-fiction cases in which a futuristic brain scan reveals to us exactly which sentences we accept by, as it were, discovering our brain's wiring diagram. Arranging it that, at a particular time, we know a Gödel code[1] of the set of arithmetical sentences we are willing to accept at that time is not entirely straightforward, even if the have the full medical resources of the starship *Enterprise* at our disposal. The trouble is that our arithmetical beliefs aren't entire the product of *a priori* cogitation. Some of our arithmetical beliefs we have because of sensory experiences we've had, for example, hearing a lecture or reading a book, and we may expect that our beliefs will change in the future on the basis of new sensory experiences. In particular, if a computer analysis of a brain snapshot taken at time t_0 tells me that the set of my arithmetical sentences I accept has Gödel code k, I might respond to this knowledge by accepting that the set of sentences with Gödel code k is consistent. But this won't mean that I believe a set of sentences that implies its own consistency. Instead, the set of my arithmetical beliefs at a later time t_1, after I've seen the computer printout, includes the statement that the set of sentence I

[1] We may that the Gödel code of a recursively enumerable set to be the Gödel number of a Σ formula that has the set as its extension.

believed at an earlier time t_0 is consistent. To arrange things so that, at a particular time, we have a code of the set of arithmetical sentences we are willing to accept at that time, we have to be more devious.

A complete circuit analysis of your brain ought to tell you, not only what sentences you accept, but what sentences you would accept in response to various sensory inputs you might have in the future, in the same way that a complete circuit analysis of an adding machine would tell you how the machine would respond to various possible combinations of keystrokes. It will provide you with a detailed description of a function f, such that, for any possible sensory input i, f(i) is a Gödel code of the set of arithmetical sentences you will accept after experiencing i. Call this function f your *transition function.*[2]

There's no telling how new sensory inputs will affect your mental states. Perhaps the song of a mockingbird will awaken an erstwhile dozing corner of your brain an enable you to enjoy a mathematical insight that would otherwise have been unavailable. We are trying to show that, if the clockwork model of the mind is correct, it is theoretically possible, with enough effort, resourcefulness, and technical firepower, for you to put yourself in a position in which you know a Gödel code for the set of arithmetical sentences you are willing to accept. To that end, let's try to minimize the effects of random disturbances like mockingbird calls, and focus our attention on the question how to accommodate the fact that, typically, learning your mental state is going to change your mental state. So suppose that you are put into a sensory deprivation chamber, where the only sensory stimulation you are in a position to enjoy is an Arabic numeral,

[2] To keep things tolerably simple, I am supposing that your brain is deterministic. In real life, the jury is still out on this, or so I understand.

which will appear on a LED display. Using the number k as a code for the state in which the numeral for k appears on the LED display, we can treat your transition function as having numerical inputs. As you enter the sensory deprivation chamber, a snapshot of your brain is taken and fed into a computer, which determines the Gödel code of your transition function. The Arabic numeral for this code is then flashed on the LED display. Let's say the number is k. Then, after you see the display, you'll know the Gödel number of the set of arithmetical sentences you are willing to accept. It's the output of the function with Gödel code k – call it "f" – on input k. You'll know your own Gödel number.

Now that you know that you are willing to accept all the members of the set coded by f(k), should you also be willing to accept that all the members of the set are true? For each sentence in the set, you have good enough reason to accept the sentence, so you have good enough reason to accept that the sentence is true. But this doesn't show that you have reason to accept that all the sentences in the set are true, since we sometimes find ourselves in a position to accept each instance of a generalization without being in a position to accept the generalization. If you are in a position to accept the thesis that all the members of f(k) are true, the evidence that leads you to accept it will be psychological, not mathematical. Observing the LED didn't give you any new mathematical insights. What it told you, assuming you trust the instruments, is something about your psychological states. The question is, now that you know the psychological fact that you are willing to accept all the outputs of f(k), does this give you good reason to accept that all the outputs of f(k) are true?

I can imagine you reasoning like this, "I know that I am careful, methodical, and clever, and that I don't accept things without good reason. I'm not the sort of person who makes mathematical mistakes. So I can be confident that the things I am willing to accept are true." I hope you don't reason this way, because to me it sounds terribly arrogant. And, indeed, such arrogance is promptly punished. If you accept that a the things you are willing to accept are true, you'll be willing to accept that the things you are willing to accept are consistent, and this means, by the Second Incompleteness Theorem, that the things you are willing to accept are inconsistent.

You might try to overcome the accusation of arrogance by adopting higher standards of acceptance, taking advantage of the fact that "willing to accept" is a vague term. "Most of the time," you will admit, "I am as error prone as the next fellow. But right now, when I talk about 'accepting' an arithmetical statement, I mean that I am willing to embrace the statement under the very highest standards of meticulous mathematical rigor. That's how I'm using the word 'accept' here, and I have programmed the transition-function-recognition program to reflect this high standard of rigor. Surely, if I have such elevated epistemic criteria, I can be confident that the things I am willing to accept are true." The trouble is that, while adopting these very high standards does indeed make it more reasonable to believe that they things you are willing to accept are true, this belief in your own veracity, while reasonable, isn't secure enough to pass your very high standards. Raising standards makes it harder for your beliefs to count as "accepted," including your belief that the things you accept are true.

In the end, I think the moral to be drawn from the Lucas-Penrose argument is a lesson in humility. We human beings are highly fallible, and we can't be sure, even when we're reasoning carefully, that the things we accept are true, or even consistent.

Let us return our attention to the case in which our theory Γ is something like Peano arithmetic, that we consciously accept. That means that we not only suppose that the members of Γ are consistent; we believe that they're true. Of course, the statement that the members of Γ are true, unlike the statement that Γ is consistent, isn't something that we are able to express within the language of arithmetic. But our belief that Γ is true will have repercussions for what purely arithmetical sentences we are willing to accept. For a given arithmetical sentence ϕ, we might or might not know whether ϕ is true, but we will know, because we regard all the members of Γ as true and we recognize that all the consequences of a true theory are true, that, if ϕ is a consequence of Γ, ϕ is true. That is, we are willing to accept all of the so-called Reflection Axioms:

$$(\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]) \rightarrow \phi).$$

The Reflection Axioms are called that because we get them, not directly from Γ, but by reflecting on the fact that Γ is a theory that we are willing to acknowledge as true.

The Reflection Axioms are statements we are willing to accept in virtue of our conscious willingness to accept Γ. Which of them are actually consequences of Γ? Well, of course, if ϕ is a consequence of Γ, then the conditional $(\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]) \rightarrow \phi)$ is a consequence of Γ, since you can prove a conditional by proving its consequent. It turns out that these are the only Reflection Axioms that are provable in Γ:

> **Löb's Theorem.** If Γ is a recursive set of sentences that includes PA, the Reflection Axiom $(\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]) \rightarrow \phi)$ is a consequence of Γ only if ϕ is a consequence of Γ.

We can derive the Second Incompleteness Theorem as a special case of Löb's Theorem by setting ϕ equal to "$\sim 0 = 0$." Conversely, we can derive Löb's Theorem from the Second Incompleteness Theorem. That wasn't the way Löb proved the theorem initially; it's a later proof, due to Saul Kripke.

Proof (Kripke): The key fact we use is that, for any ψ and θ, $\mathrm{Bew}_\Gamma([\ulcorner(\psi \rightarrow \theta)\urcorner])$ is provably equivalent to $\mathrm{Bew}_{\Gamma\cup\{\psi\}}([\ulcorner\theta\urcorner])$. This is easy to see. If we have a derivation of $(\psi \rightarrow \theta)$ from Γ, we can get a derivation of θ from $\Gamma \cup \{\psi\}$ by adding ψ as a line by Premise Introduction, then adding θ by *Modus Ponens*. If, conversely, we have a proof of θ from $\Gamma \cup \{\psi\}$, we get a derivation of $(\psi \rightarrow \theta)$ from Γ by Conditional Proof.

We prove Löb's theorem by proving its contrapositive, that is, by assuming that ϕ isn't a consequence of Γ and deriving that consequence that $(\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]) \rightarrow \phi)$ isn't a consequence of Γ. Since ϕ isn't a consequence of Γ, $\Gamma \cup \{\sim \phi\}$ is consistent. It follows from the Second Incompleteness theorem that $\Gamma \cup \{\sim \phi\}$ doesn't prove its own consistency, so that we have:

$$\Gamma \cup \{\sim \phi\} \nvdash \mathrm{Con}(\Gamma \cup \{\sim \phi\}).$$

$$\Gamma \cup \{\sim \phi\} \nvdash \sim \mathrm{Bew}_{\Gamma\cup\{\sim\phi\}}([\ulcorner \sim 0 = 0\urcorner]).$$

$$\Gamma \cup \{\sim \phi\} \nvdash \sim \mathrm{Bew}_\Gamma([\ulcorner(\sim \phi \rightarrow \sim 0 = 0)\urcorner]).$$

$$\Gamma \cup \{\sim \phi\} \nvdash \sim \mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner])$$

(because $(\sim \phi \rightarrow \sim 0 = 0)$ and ϕ are provably equivalent).

$$\Gamma \nvdash (\sim \phi \rightarrow \sim \mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner])).$$

$$\Gamma \nvdash (\mathrm{Bew}_\Gamma([\ulcorner\phi\urcorner]) \rightarrow \phi).\boxtimes$$

The way Löb's Theorem came about was this: Gödel constructed a sentence that asserted its own unprovability in PA, and he showed that that sentence was true but unprovable in PA.

Leon Henkin was curious what would happen if, instead, you constructed a sentence that asserted its own provability in PA. Would such a sentence be provable? Would it be true? Löb answered Henkin's question by showing that, if ϕ is a sentence that asserts its own provability in PA, so that

$$\text{PA} \vdash (\phi \leftrightarrow \text{Bew}_{\Gamma}([\ulcorner \phi \urcorner])),$$

then ϕ is provable in PA (and so true). Löb sent his paper to the *Journal of Symbolic Logic*, which sent it to Henkin to referee. Henkin noticed that Löb's proof only used the right-to-left direction of the hypothesis. Thus Löb's Theorem, as we have it today, was born. Here is Löb's proof:

Proof (Löb): Given ϕ, use the Self-Referential Lemma to construct a sentence δ such that:

(i) $\Gamma \vdash (\delta \leftrightarrow (\text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \rightarrow \phi))$.

(i) yields this, by sentential calculus:

(ii) $\Gamma \vdash (\delta \rightarrow (\text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \rightarrow \phi))$.

(L1) gives us this:

(iii) $\Gamma \vdash \text{Bew}_{\Gamma}([\ulcorner (\delta \rightarrow (\text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \rightarrow \phi)) \urcorner])$.

Two applications of (L3) give us this:

(iv) $\Gamma \vdash (\text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \rightarrow (\text{Bew}_{\Gamma}([\ulcorner \text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \urcorner]) \rightarrow \text{Bew}_{\Gamma}([\ulcorner \phi \urcorner])))$.

(L2) yields this:

(v) $\Gamma \vdash (\text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \rightarrow \text{Bew}_{\Gamma}([\ulcorner \text{Bew}_{\Gamma}([\ulcorner \delta \urcorner]) \urcorner]))$.

By a truth-functional inference of the form

$(\alpha \rightarrow (\beta \rightarrow \gamma))$

$(\alpha \rightarrow \beta)$

$\therefore (\alpha \rightarrow \gamma)$

we derive:

(vi) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\delta\urcorner]) \rightarrow \mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner]))$.

Assume, as the hypothesis of Löb's Theorem:

(vii) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\phi\urcorner]) \rightarrow \phi)$.

(vi) and (vii) give us this:

(viii) $\Gamma \vdash (\mathrm{Bew}_{\Gamma}([\ulcorner\delta\urcorner]) \rightarrow \phi)$

(i) and (viii) give us:

(ix) $\Gamma \vdash \delta$.

Using (L1), we derive:

(x) $\Gamma \vdash \mathrm{Bew}_{\Gamma}([\ulcorner\delta\urcorner])$.

(viii) and (x) give us:

(xi) $\Gamma \vdash \phi$,

as required.⊠